HAL
archives-ouvertes.fr

Learning Multi-Modal Dictionaries: Application to Audiovisual Data

Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhé, Sylvain Lesage, Rémi Gribonval

▶ **To cite this version:**

Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhé, Sylvain Lesage, et al.. Learning Multi-Modal Dictionaries: Application to Audiovisual Data. Proc. of International Workshop on Multimedia Content Representation, Classification and Security (MCRCS'06), Sep 2006, Istanbul, Turkey. Springer-Verlag, 4105, pp.538–545, 2006, LNCS. <10.1007/11848035_71>. <inria-00544773>

HAL Id: inria-00544773
https://hal.inria.fr/inria-00544773

Submitted on 8 Feb 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

Learning Multi-Modal Dictionaries: Application to Audiovisual Data

Gianluca Monaci[1], Philippe Jost[1], Pierre Vandergheynst[1], Boris Mailhe[2], Sylvain Lesage[2], and Rémi Gribonval[2]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Institute, CH-1015 Lausanne, Switzerland
{gianluca.monaci,philippe.jost,pierre.vandergheynst}@epfl.ch
[2] IRISA-INRIA, Campus de Beaulieu, 35042 Rennes CEDEX, France
{boris.mailhe,sylvain.lesage,remi.gribonval}@irisa.fr

Abstract. This paper presents a methodology for extracting meaningful synchronous structures from multi-modal signals. Simultaneous processing of multi-modal data can reveal information that is unavailable when handling the sources separately. However, in natural high-dimensional data, the statistical dependencies between modalities are, most of the time, not obvious. Learning fundamental multi-modal patterns is an alternative to classical statistical methods. Typically, recurrent patterns are shift invariant, thus the learning should try to find the best matching filters. We present a new algorithm for iteratively learning multi-modal generating functions that can be shifted at all positions in the signal. The proposed algorithm is applied to audiovisual sequences and it demonstrates to be able to discover underlying structures in the data.

1 Introduction

Multi-modal signal analysis has received an increased interest in the last years. Multi-modal signals are sets of heterogeneous signals originating from the same phenomenon but captured using different sensors, having thus different characteristics. Each modality typically brings some information about the others and their simultaneous processing can uncover relationships that are otherwise unavailable when considering the signals separately. In this work we analyze a broad class of multi-modal signals exhibiting correlations along time. Mutual dependencies along time can be discovered observing the temporal evolution of the signals. Examples come from neuroscience, where EEG and functional MRI (fMRI) are jointly analyzed to study brain activation patterns [1]; environmental science, where different spatio-temporal measurements are correlated to discover connections between local and global phenomena [2]; multimedia signal processing, where audio and video sequences are combined to localize the sound source in the video [3–6]. Also humans exploit the temporal co-occurrence of acoustic and visual stimuli to enhance their comprehension of audiovisual scenes [7].

Temporal correlation across modalities is exploited by seeking for patterns showing a certain degree of synchrony. Typically, research efforts have focused

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

on the statistical modelling of the dependencies between modalities. In [1] EEG and fMRI structures having maximal temporal covariance are extracted, in [3] projections onto maximally independent audiovisual subspaces are considered, in [4] the video components correlated with the audio are detected maximizing the Mutual Information between audio energy and pixel values, in [5] audio-video quantities are correlated using Canonical Correlation Analysis. However, the features employed to represent the different modalities are basic and barely connected with the physics of the observed phenomena (e.g. video sequences are represented using time series of pixel intensities). This can be a limit of existing approaches: multi-modal features having low structural content can be difficult to extract and manipulate. Moreover, the interpretation of the results can be problematic without an accurate modelling of the observed phenomenon.

However, the problem can be attacked from another point of view. The complexity of multi-modal fusion algorithms can be concentrated on the modelling of the modalities, so that *meaningful* structures can be extracted from the signals and synchronous patterns can be easily detected.

An application of this paradigm can be found in [6], where *meaningful* audio-visual structures are defined as temporally proximal audio-video *events*. Audio and video signals are represented in terms of their most salient structures over redundant dictionaries of functions, making possible the definition of audio-video events. The synchrony of these events appears to reflect the presence of a common source, which is effectively localized. The key idea of this approach is to use high-level features to represent signals, which are introduced by making use of codebooks of functions. The audio signal is decomposed as a sum of Gabor atoms, while the video sequence is expressed as a combination of edge-like functions which are tracked through time. Such audio and video representations are still quite general, and can be employed to represent any audiovisual sequence.

However, the main advantage of dictionary-based techniques is the freedom in designing the dictionary, which can be efficiently tailored to closely match signal structures [8–12]. Often, natural signals have highly complex underlying structures, which makes it difficult to explicitly define a link between a class of signals and a dictionary. This paper presents a learning algorithm that tries to capture the underlying structures of multi-modal signals enforcing synchrony between modalities. We propose to learn *multi-modal generating functions*, i.e. functions constituted of multiple components, one for each signal modality, and that exist in the same time slot. Each function defines a set of atoms corresponding to all its translations. This is notably motivated by the fact that natural signals often exhibit statistical properties invariant to translation, and the use of generating functions allows to generate big dictionaries while using only few parameters. The proposed algorithm learns the generating functions successively and can be stopped when a sufficient number of atoms have been found.

The algorithm presented in this paper is the generalization to multi-modal signals of the MoTIF algorithm [13]. Following this work, in the next section we reformulate the problem of learning multi-modal generating functions.

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

2 Learning Multi-Modal Dictionaries

Formally, the aim is to learn a collection $\mathcal{G} = \{g_k\}_{k=1}^{K}$ of multi-component generating functions g_k such that a highly redundant dictionary $\mathcal{D}$ adapted to a class of signals can be created by applying all possible translations to the generating functions of $\mathcal{G}$. The function g_k can consist in an arbitrary number of components. For simplicity, we will treat here the bimodal case; however, the extension to the multichannel case is straightforward. A 2-components generating function can be written as $g_k = (g_k^{(1)}, g_k^{(2)})$, where $g_k^{(1)}$ and $g_k^{(2)}$ are the components of g_k on the two modalities. The components do not have to be homogeneous in dimensionality; however, they have to share a common temporal dimension.

For the rest of the paper, we assume that the signals denoted by lower case letters are discrete and of infinite size. Finite size vectors and matrices are denoted with bold characters. Let $T_p^{(i)}$ be the operator that translates an infinite signal on channel i by $p \in \mathbb{Z}$ samples. Let the set $\{T_p^{(i)} g_k^{(i)}\}$ contain all possible atoms generated by applying the translation operator to $g_k^{(i)}$. The dictionary generated by $\mathcal{G}$ is $\mathcal{D} = \{\{(T_p^{(1)} g_k^{(1)}, T_p^{(2)} g_k^{(2)})\}, k = 1 \dots K\}$. The couple of operators $(T_p^{(1)}, T_p^{(2)})$ translates the signals synchronously on the two channels, in such a way that their temporal proximity is preserved.

The learning is done using a training set of N bimodal signals $\{(f_n^{(1)}, f_n^{(2)})\}_{n=1}^{N}$, where $f_n^{(1)}$ and $f_n^{(2)}$ are the components of the signal on the two modalities. The signals have infinite size and they are non null on their support of size $(S_{f^{(1)}}, S_{f^{(2)}})$. Similarly, the size of the support of the generating functions to learn is $(S_{g^{(1)}}, S_{g^{(2)}})$ such that $S_{g^{(1)}} < S_{f^{(1)}}$ and $S_{g^{(2)}} < S_{f^{(2)}}$. The proposed algorithm learns translation invariant filters iteratively. For the first one, the aim is to find g_1 such that the dictionary $\{(T_p^{(1)} g_1^{(1)}, T_p^{(2)} g_1^{(2)})\}$ is the most correlated in mean with the signals in the training set. Hence, it is equivalent to the following optimization problem:

$$\text{UP} : g_1^{(i)} = \underset{\|g^{(i)}\|_2=1}{\arg\max} \sum_{n=1}^{N} \max_{p_n} | \langle f_n^{(i)}, T_{p_n}^{(i)} g^{(i)} \rangle |^2 , \tag{1}$$

which has to be solved simultaneously for the two modalities ($i = 1, 2$).

For learning the successive generating functions, the problem can be slightly modified to include a constraint penalizing a generating function if a similar one has already been found. Assuming that $k-1$ generating functions have been learnt, the optimization problem to find g_k can be written as:

$$\text{CP} : g_k^{(i)} = \underset{\|g^{(i)}\|_2=1}{\arg\max} \frac{\sum_{n=1}^{N} \max_{p_n} | \langle f_n^{(i)}, T_{p_n}^{(i)} g^{(i)} \rangle |^2}{\sum_{l=0}^{k-1} \sum_p | \langle g_l^{(i)}, T_p^{(i)} g^{(i)} \rangle |^2} , \tag{2}$$

which again has to be solved simultaneously for the two modalities ($i = 1, 2$).

Finding the best solution to the unconstrained problem (UP) or the constrained problem (CP) is hard. However, the problem can be split into several

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

simpler steps following a *localize and learn* paradigm [13]. Such strategy is particularly suitable for this scenario, since we want to learn *meaningful* synchronous patterns that are localized in time and that represent well the signals. Thus, we propose to perform the learning by iteratively solving the following four steps:

1. **(*localize*)** for a given generating function $g_k^{(1)}[j]$ at iteration j, find the best translations $p_n[j]$,
2. **(*learn*)** update $g_k^{(2)}[j]$ by solving UP (1) or CP (2), where the optimal translations p_n are fixed to the previous values $p_n[j]$,
3. **(*localize*)** find the best translations $p_n[j+1]$ using the function $g_k^{(2)}[j+1]$,
4. **(*learn*)** update $g_k^{(1)}[j+1]$ by solving UP (1) or CP (2), where the optimal translations p_n are fixed to the previous values $p_n[j+1]$.

Note that the temporal synchrony between generating functions on the two channels is simply enforced at the learning steps (2 and 4), where the optimal translation p_n found for one modality is also kept for the other one.

The first and third steps consist in finding the location of the maximum correlation between each learning signal $f_n^{(i)}$ and the generating function $g^{(i)}$.

Let now consider the second and fourth steps and define $\mathbf{g}_k^{(i)} \in \mathbb{R}^{S_{g^{(i)}}}$ the restriction of the infinite size signal $g_k^{(i)}$ to its support. As the translation admits a well defined adjoint operator, $\langle f_n^{(i)}, T_{p_n}^{(i)} g_k^{(i)} \rangle$ can be replaced by $\langle T_{-p_n}^{(i)} f_n^{(i)}, g_k^{(i)} \rangle$. Let $\mathbf{F}^{(i)}[j]$ be the matrix ($S_{f^{(i)}}$ rows, N columns), whose columns are made of the signals $f_n^{(i)}$ shifted by $-p_n[j]$. More precisely, the j^{th} column of $\mathbf{F}^{(i)}[j]$ is $\mathbf{f}_{n,-p_n[j]}^{(i)}$, the restriction of $T_{-p_n[j]}^{(i)} f_n^{(i)}$ to the support of $g_k^{(i)}$, of size $S_g^{(i)}$. We denote $\mathbf{A}^{(i)}[j] = \mathbf{F}^{(i)}[j]\, \mathbf{F}^{(i)}[j]^T$.

With these notations, the second step of the *unconstrained* problem can be written:

$$\mathbf{g}_k^{(i)}[j+1] = \underset{||\mathbf{g}^{(i)}||_2=1}{\arg\max}\ {\mathbf{g}^{(i)}}^T \mathbf{A}^{(i)}[j]\, \mathbf{g}^{(i)} \tag{3}$$

where $.^T$ denotes the transposition. The best generating function $\mathbf{g}_k^{(i)}[j+1]$ is the eigenvector associated with the biggest eigenvalue of $\mathbf{A}^{(i)}[j]$.

For the *constrained* problem, we want to force $g_k^{(i)}[j+1]$ to be as de-correlated as possible from all the atoms in $\mathcal{D}_{k-1}$. This corresponds to minimizing

$$\sum_{l=1}^{k-1} \sum_{p} |\langle T_{-p} g_l^{(i)}, g^{(i)} \rangle|^2 \tag{4}$$

or, denoting

$$\mathbf{B}_k^{(i)} = \sum_{l=1}^{k-1} \sum_{p} \mathbf{g}_{l,-p}^{(i)}\, {\mathbf{g}_{l,-p}^{(i)}}^T, \tag{5}$$

to minimizing ${\mathbf{g}^{(i)}}^T \mathbf{B}_k^{(i)} \mathbf{g}^{(i)}$. With these notations, the constrained problem can be written as:

$$\mathbf{g}_k^{(i)}[j+1] = \underset{||\mathbf{g}^{(i)}||_2=1}{\arg\max}\ \frac{{\mathbf{g}^{(i)}}^T \mathbf{A}^{(i)}[j]\, \mathbf{g}^{(i)}}{{\mathbf{g}^{(i)}}^T \mathbf{B}_k^{(i)} \mathbf{g}^{(i)}} \tag{6}$$

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

The best generating function $\mathbf{g}_k^{(i)}[j+1]$ is the eigenvector associated to the biggest eigenvalue of the generalized eigenvalue problem defined in (6). Defining $\mathbf{B}_1^{(i)} = \mathbf{Id}$, we can use CP for learning the first generating function $\mathbf{g}_1$.

The unconstrained single-channel algorithm has been proven to converge in a finite number of iterations to a generating function locally maximizing the unconstrained problem [13]. We observed on numerous experiments that the constrained algorithm and the multi-modal constrained algorithm typically converge in few steps to a stable solution independently of the initialization.

3 Experiments

In the first experiment a multi-modal dictionary is learned on a set of three audiovisual sequences representing the same mouth uttering the digits from zero to nine in English. In this case the two modalities are audio and video, which share a common temporal axis. The audio was recorded at 44 kHz and it was sub-sampled to 8 kHz, while the video was recorded at 29.97 frames/second (fps) and at a resolution of 70×110 pixels. The total length of the training sequences is 806 frames, i.e. approximately 27 seconds. Note that the sampling frequencies along the time axis for the two modalities are different, thus when passing from one modality to the other a re-sampling factor r equal to the ratio between the two frequencies has to be applied, i.e. $r = 8000/29.97 \approx 267$. The audio signal is considered as is while the video is whitened using the procedure described in [12] to speed up the training. The learning is performed on audio-video patches extracted from the original signals. We use patches whose size $f_n^{(a)}$ is 6407 audio samples, while $f_n^{(v)}$ is 31×31 pixels in space and 23 frames in time. We learn 20 generating functions consisting of an audio component of 3204 samples and a video component of size 16×16 pixels in space and 12 frames in time. The first 15 elements of the learned dictionary are shown in Fig. 1. The video component of each function is shown on the left, with time proceeding left to right, while the audio part is on the right, with time on the horizontal axis.

Concerning the video components, they are spatially localized and oriented edge detector functions. They oscillate in time, describing typical movements of different parts of the mouth during the utterances. The audio parts of the generating functions contain almost all the numbers present in the training sequences. In particular, when listening to the waveforms, one can clearly distinguish the words *zero* (functions #1, #8, #11, #14), *one* (#4), *two* (#3, #10), *three* (#5), *five* (#7, #15), *seven* (#12), *nine* (#6, #9, #13). Typically, different instances of the same number have different characteristics, like length or frequency content (i.e. compare audio functions #1, #8, #11 and #14). As already observed in [13], both components of generating function #2 are mainly high frequency due to the de-correlation constraint with the first atom.

In order to study the differences between the single-channel learning algorithm and its multi-modal version, we learn 20 audio generating functions on the audio training set used in the previous experiment using the single-channel MoTIF algorithm [13]. The resulting learned dictionary has characteristics similar

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

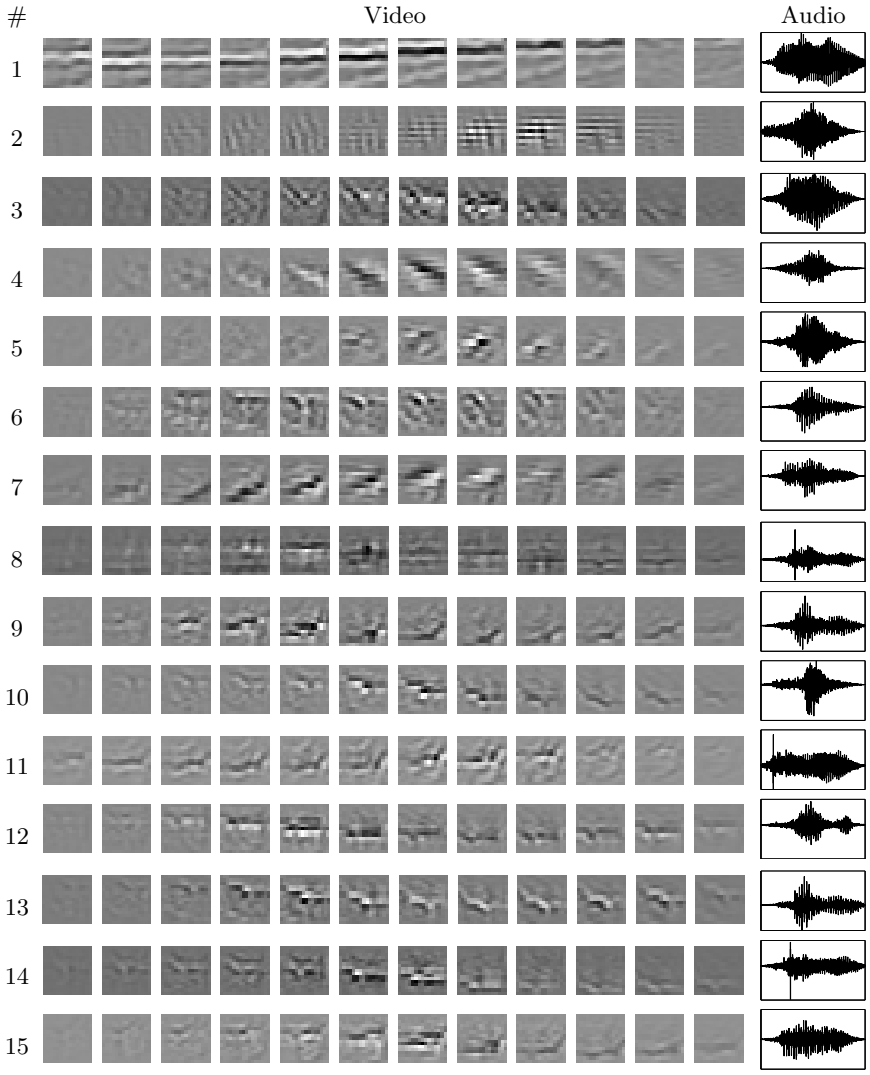

Fig. 1. Audio-video generating functions. Shown are the first 15 functions learned, each consisting on an audio and a video component. Video components are on the left, with time proceeding left to right. Audio components are on the right, with time on the horizontal axis.

to those of the audio components of the multi-modal dictionary shown in Fig. 1. However, the variety of functions that are learned is smaller than in the previous experiment. In particular, when listening to the waveforms it is possible to distinguish instances of the words *zero*, *one*, *three*, *five*, *nine*. Interestingly, it seems that forcing the algorithm to learn meaningful audio features synchronous to meaningful video features allows to uncover richer and wider-ranging structures.

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

Fig. 2. Sample frame of the test audiovisual sequence. The white cross highlights the median spatial position of the video maxima.

In the third experiment we test how the learned dictionary is able to recover meaningful audiovisual patterns in real multimedia sequences. We consider a clip consisting in two persons in front of the camera arranged as in Fig. 2. One of the subjects (the person on the left) is uttering digits in English, while the other one is mouthing *exactly the same words*. The clip can be downloaded through `http://lts2www.epfl.ch/~monaci/avLearn.html`. The audio track is at 8 kHz, while the video is at 29.97 fps and at a resolution of 480×720 pixels. The speaker is the same subject whose mouth was used to train the multi-modal dictionary in Fig. 1; however, the training sequences are different from the test sequence. We want to underline that such a sequence is particularly difficult to analyze, since both persons are mouthing the same words at the same time. The task of associating the sound with the "real" speaker is thus non-trivial.

The audio track of the test clip is filtered with each audio component of the 20 learned generating functions. For each audio function we keep the time position of maximum projection and we consider a window of 31 frames around this time position in the video. This video patch is filtered with the corresponding video component and the spatio-temporal position of maximum projection between the video and the learned video generating function is kept. Thus, for each multi-modal function we obtain the position of maximal projection over the time axis for the audio part and the location of maximal projection over the image plane and over time for the video component. What we expect is that the spatial position of the video maxima are localized on the speaker's mouth and that the relative shift between the time positions of the audio and video maxima is small. The mean shift between audio-video pairs is found to be equal to 1.5 frames, which is a reasonably good result considering the errors introduced by the re-sampling applied to audio-video signals. The median spatial position of the video maxima is located on the speaker's mouth, as shown in Fig. 2. In this case the median is consider in order to filter out spurious erroneous maxima positions that would bias the centroid estimate. Using the learned dictionary it is possible to detect synchronous audio-video patterns, recovering the synchrony between audio and video tracks and localizing the sound source on the video sequence.

© 2006 Springer. This is the author version of an article published in Springer Lecture Notes in Computer Sciences. The original publication is available at www.springerlink.com

4 Conclusions

In this paper we present a new method to learn translation invariant multi-modal functions adapted to a class of multi-component signals. Generating waveforms are iteratively found using a *localize and learn* paradigm which enforces temporal synchrony between modalities. A constraint in the objective function forces the learned waveforms to have low correlation, such that no function is picked several times. The algorithm seems to capture well the underlying structures in the data. The dictionary includes elements that describe typical audiovisual features present in the training signals. The learned functions have been used to analyze a complex sequence, obtaining encouraging results in recovering audio-video synchrony and localizing the sound source on the video sequence. Applications of this technique to other types of multi-modal signals, like climatologic or EEG-fMRI data, are foreseen.

References

1. Martínez-Montes, E., Valdés-Sosa, P.A., Miwakeichi, F., Goldman, R.I., Cohen, M.S.: Concurrent EEG/fMRI analysis by multiway partial least squares. Neuroimage **22** (2004) 1023–1034
2. Carmona-Moreno, C., Belward, A., Malingreau, J., Garcia-Alegre, M., Hartley, A., Antonovskiy, M., Buchshtaber, V., Pivovarov, V.: Characterizing inter-annual variations in global fire calendar using data from earth observing satellites. Global Change Biology **11** (2005) 1537–1555
3. Smaragdis, P., Casey, M.: Audio/visual independent components. In: Proc. of ICA. (2003) 709–714
4. Fisher III, J.W., Darrell, T.: Speaker association with signal-level audiovisual fusion. IEEE Transactions on Multimedia **6** (2004) 406–413
5. Kidron, E., Schechner, Y., Elad, M.: Pixels that sound. In: CVPR. (2005) 88–95
6. Monaci, G., Divorra Escoda, O., Vandergheynst, P.: Analysis of multimodal sequences using geometric video representations. Signal Processing **in press** (2006) [Online] Available: `http://lts2www.epfl.ch/`.
7. Driver, J.: Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature **381** (1996) 66–68
8. Bell, A., Sejnowski, T.: The "independent components" of natural scenes are edge filters. Vision research **37** (1997) 3327–3338
9. Lewicki, M., Sejnowski, T.: Learning overcomplete representations. Neural computation **12** (2000) 337–365
10. Abdallah, S., Plumbley, M.: If edges are the independent components of natural images, what are the independent components of natural sounds? In: Proc. of ICA. (2001) 534–539
11. Kreutz-Delgado, K., Murray, J., Rao, B., Engan, K., Lee, T., Sejnowski, T.: Dictionary learning algorithms for sparse representation. Neural Computation **15** (2003) 349–396
12. Olshausen, B.: Learning sparse, overcomplete representations of time-varying natural images. In: Proc. of ICIP. (2003)
13. Jost, P., Vandergheynst, P., Lesage, S., Gribonval, R.: MoTIF: an efficient algorithm for learning translation invariant dictionaries. In: Proc. of ICASSP. (2006)