ANALYSIS OF MULTIMODAL SIGNALS USING REDUNDANT REPRESENTATIONS

Gianluca Monaci, Oscar Divorra Escoda, Pierre Vandergheynst

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute
CH-1015 Lausanne, Switzerland
e-mail: {gianluca.monaci, oscar.divorra, pierre.vandergheynst}@epfl.ch

ABSTRACT

In this work we explore the potentialities of a framework for the representation of audio-visual signals using decompositions on overcomplete dictionaries. Redundant decompositions may describe audio-visual sequences in a concise fashion, preserving good representation properties thanks to the use of redundant, well designed, dictionaries. We expect that this will help us overcome two typical problems of multimodal fusion algorithms. On one hand, classical representation techniques, like pixel-based measures (for the video) or Fourier-like transforms (for the audio), take into account only marginally the physics of the problem. On the other hand, the input signals have large dimensionality. The results we obtain by making use of sparse decompositions of audio-visual signals over redundant codebooks are encouraging and show the potentialities of the proposed approach to multimodal signal representation.

1. INTRODUCTION

The problem we are studying in this work is that of correlating audio tracks with visual data to detect those regions in an image sequence from which the sound is originated. The topic was first faced by Hershey and Movellan [1]. They measured the correlation between audio and video using an estimate of the mutual information derived from the Pearson correlation coefficient between the energy of an audio track and the value of single pixels. Slaney and Covell [2] generalize this approach and look for a method able to measure the synchrony between audio signals and video facial images. In order to deduce a relationship between the cepstral representation of the audio and the video pixels, the authors use canonical correlation analysis. In [3], an approach based on Markov chains modeling audio and video signals is proposed. The audio-visual consistency is assessed by maximizing the mutual information between the power spectrum coefficients of the audio and the video pixels intensity change. For the three methods, audio and video joint densities are deduced by training on audio-video sequences. A method that does not make use of any previous model training is that proposed by Fisher and Darrell [4]. The algorithm is based on a probabilistic generation model that is used to learn audio and video linear features that maximize the mutual information between the different modalities. A slightly different approach is used in [5], where a methodology for extracting audio-visual independent components from video streams is presented. However, this technique is not able to deal with dynamic scenes.

In this paper, we explore a completely new framework for the representation of multimodal signal in the context of audio-visual fusion. This is based on the sparse decomposition of signals over atoms dictionaries using Matching Pursuits [6] (MP). An appropriate decomposition of a signal over a well designed redundant dictionary provides an interpretation of the information in terms of the most salient signal structures. By representing the video using image structures (*atoms*) that evolve in time, in fact, we deal with dynamic features that have a true geometrical meaning, that is not the case when using pixel-based representations. At the same time, the MP decomposition provides a sparse representation of the information, allowing a considerable reduction of the dimensionality of the input signals. This should allow us to handle information in an easier and faster way, and thus to develop relatively simple and intuitive, but effective, fusion criteria. In order to combine audio and video representations, we use a "classical" measure of correlation, the *Pearson* correlation coefficient. The obtained results show that our technique allows to locate those image structures from which the audio signals are originate.

2. AUDIO AND VIDEO REPRESENTATIONS

Audio and video signals are represented using MP decompositions over redundant dictionaries. In the next sections, we will present the MP algorithm for 1-D signals, and the techniques that have been developed to extend it to the complex case of video sequences.

The authors acknowledge the support of the Swiss National Science Foundation through the IM.2 National Center of Competence for Research.
Web page: http://lts2www.epfl.ch.

2.1. Audio Decomposition

The audio signal $a(t)$ is decomposed using the MP algorithm over a redundant dictionary $\mathcal{D}_{\mathcal{A}}$ of unit norm functions called *atoms*. The family of atoms that form $\mathcal{D}_{\mathcal{A}}$ is generated by scaling by s, translating in time by u and modulating in frequency by ξ a generating function $g(t) \in L^2(R)$. Indicating with an index γ the set of transformations (s, u, ξ), an atom can be expressed as

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\Big(\frac{t-u}{s}\Big) e^{i\xi t}\,. \tag{1}$$

In our case, we consider a dictionary of Gabor atoms. That is, the generating function $g(t)$ is a normalized Gaussian window, which has been chosen for its optimal time-frequency localization [7].

The first step of the MP algorithm decomposes a as

$$a = \langle a, g_{\gamma_0}\rangle g_{\gamma_0} + R^1 a\,, \tag{2}$$

where $R^1 a$ is the residual component after approximating a in the subspace described by g_{γ_0}. The function g_{γ_0} is chosen such that the projection $|\langle a, g_{\gamma_0}\rangle|$ is maximal. This procedure is recursively applied, and after N iterations the signal a is represented as

$$a = \sum_{n=0}^{N-1} \langle R^n a, g_{\gamma_n}\rangle g_{\gamma_n} + R^N a\,, \tag{3}$$

where $R^0 = a$ and $R^n a$ is the residual after n iterations.

2.2. Video Decomposition

The image sequence is represented using the MP algorithm proposed by Divorra and Vandergheynst [8]. This technique decomposes a sequence into a set of 2-D atoms evolving in time, allowing to represent salient geometric video components tracking their temporal transformations.

An iteration on the video MP algorithm decomposes the first frame of the sequence over a redundant dictionary $\mathcal{D}_{\mathcal{V}}$ of 2-D anisotropic atoms [9]:

$$I = \sum_{\gamma_i \in \Omega} c_{\gamma_i} g_{\gamma_i}\,, \tag{4}$$

where i is the summation index, c_γ corresponds to the projection coefficient for every atom g_γ and Ω is the subset of selected atom indexes from dictionary $\mathcal{D}_{\mathcal{V}}$. The changes suffered from a frame I_t to I_{t+1} are modelled as the application of an operator F_t to the image I_t such that $I_{t+1} = F_t(I_t)$ and $I_{t+1} = \sum_{\gamma_i \in \Gamma} F_t^{\gamma_i}(c_{\gamma_i}^t g_{\gamma_i}^t)$, where F_t represents the set of transformations F_t^γ of all atoms that approximate each frame. A MP-like approach similar to that used for the first frame is applied to retrieve the new set of g_γ^{t+1} (and the associated transformation F_t). At every greedy decomposition iteration only a subset of functions of the general dictionary is considered to represent each deformed atom. This subset is defined according to the past geometrical features of every atom in the previous frame, such that only a limited set of transformations are possible. The formulation of the MP approach to geometric video representation is complex and is treated in detail in [8], to which the interested reader is referred.

3. AUDIO-VISUAL FUSION

The described decompositions of audio and video sequences represent salient parameterizations of these signals. Thus, a natural and straightforward way to relate audio and video sequences is to compare these parametric representations. The considered audio-visual features are presented in Sections 3.1 and 3.2, while the criteria that are used to relate them are introduced in Section 3.3.

3.1. Audio Features

The audio representation that we obtain from the MP decomposition it is not directly exploitable. It has to be further processed in order to be easily compared with the evolution of the video parameters. Basically, we require one feature, composed of the same number of samples L as the image sequence.

In this work, we have decided to use a simple approach exploiting the properties of sparse signal representations over redundant dictionaries. The MP decomposition of the audio track, in fact, performs a denoising of the signal, pointing out the most relevant signal structures. Our audio feature is obtained by simply averaging the energy present at each time instant, where the time-frequency energy distribution of the audio signal is found by decomposing it with the MP algorithm presented in Section 2.1. Our feature is similar to those described in [3, 4], with the difference that we attribute to each frequency component the same weight. This approach is of course a very simple one, but the fine time-frequency resolution of the dictionary decomposition allows us to obtain a description that captures nicely the evolution of the audio track. We show in Fig. 1 (left) the audio feature obtained for one of the tested sequences.

3.2. Video Features

When dealing with the video signal, basically all the reviewed approaches use pixel intensities as video features, with the exception of [3], where the pixel intensity change in a 3×3 averaging spatial window is considered. Pixel-related quantities seem to us a poor source of information, that, in addition, has a huge dimensionality. Moreover, it is sensitive to noise and does not consider image structures, since spatial correlation is not exploited. We have decided, thus, to explore the possibilities offered by the greedy video decomposition technique presented in Section 2.2. In this

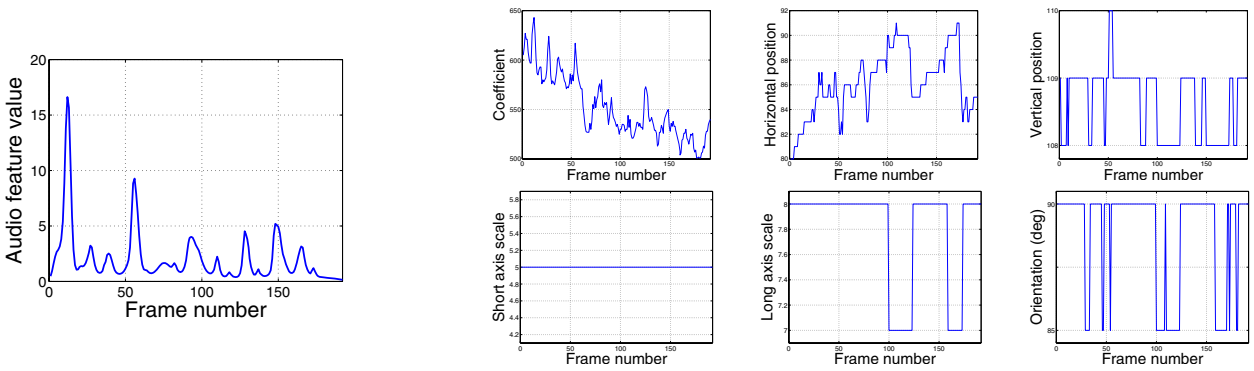

Fig. 1. Audio feature representing the acoustic signal of one tested sequences (left), and temporal evolution of the parameters of one atom used to decompose the image sequence (right). From left to right and from top to bottom: Coefficient, horizontal position, vertical position, short axis scale, long axis scale and rotation. Only the first row parameters (coefficient and positions) are considered as video features.

way, we hope to be able to track important geometric features over time and to parameterize those transformations that represent changes in the scene. The output of the MP algorithm is a set of atoms parameters that describe the temporal evolution of 3-D video features. Each atom is characterized by a coefficient, 2 position parameters, 2 scale parameters and a rotation, i.e. 6 parameters. Fig. 1 (right) shows an example of atom parameters evolution as a function of time.

The video features we consider, however, are not all these 6 variables. The scale parameters have been discarded, since they carry few information about the mouth movements. From our experiments, the atom orientation brought an unprecise description of the real geometric feature. We have thus chosen to discard it. Therefore, we have decided to employ only the atoms coefficient and positions as video features, obtaining 3 descriptors per atom. Each video sequence is represented with N time-evolving atoms. Hence, we end up with $3 \times N$ functions composed of L samples.

3.3. Fusion Criteria

Starting from the atomic representations obtained using the procedures described in Section 2, we want to detect those video atoms that are more correlated with the audio feature.

Information theoretic formulations such [3, 4] use the *Mutual Information* measure as fusion criterium [10]. However, when few data samples are available, there may be problems concerning probability density estimations. Thus, we have decided to employ a simple but robust measure of correlation, the *Pearson correlation coefficient* [11]. The *Pearson* coefficient is a parametric measure of correlation and reflects the degree of linear relationship between two variables. The observations for both variables should be approximately (bivariate) normally distributed. The *Pearson* correlation coefficient is easy and fast to compute, and allows the definition of a statistical test to estimate the significance of the considered coefficient. Given two observation vectors $\mathbf{x}$ and $\mathbf{y}$ of length n, the value of the Pearson coefficient $\hat{\rho}$ between $\mathbf{x}$ and $\mathbf{y}$ is computed as

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - E\{\mathbf{x}\})(y_i - E\{\mathbf{y}\})}{\sqrt{\sum_{i=1}^{n}(x_i - E\{\mathbf{x}\})^2}\sqrt{\sum_{i=1}^{n}(y_i - E\{\mathbf{y}\})^2}}\,, \quad (5)$$

where $E\{\mathbf{x}\}$ and $E\{\mathbf{y}\}$ denote the mean values of $\mathbf{x}$ and $\mathbf{y}$.

For each video feature, the value of the correlation $\hat{\rho}$ with the audio feature is computed. A probability p associated to the correlation coefficient is also computed, in order to assess the significance of the value of $\hat{\rho}$. When the true correlation is zero, the quantity

$$\hat{f}_t(\hat{\rho}) = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \quad (6)$$

belongs to a *Student's* distribution with $n-2$ degrees of freedom, $f_t(n-2)$, being n the number of samples [11]. If the probability p that $\hat{f}_t(\hat{\rho})$ belongs to $f_t(n-2)$ is small, then the correlation is significant. Since each atom is described by 3 time-evolving features, for each video atom we have 3 correlation values. We select those atoms for which all 3 coefficients $\hat{\rho}$ have small probability p.

4. EXPERIMENTS

The framework we have developed is used to detect the video region from which the corresponding audio signal is originated. Experiments have been carried out on video streams representing one person speaking in front of the camera. The visual data was recorded at 25 frames per second at a resolution of 144×176 pixels. The soundtrack was collected at 48 kHz and sub-sampled in order to obtain a signal at 8 kHz.

The video frames are high-pass filtered and decomposed using the algorithm described in Section 2.2, obtaining a set of 2-D time-evolving atoms. The audio part is decomposed using MP over a dictionary of Gabor atoms, using the *LastWave* implementation of MP for 1-D signals [12]. Based

(a) (b) (c)

Fig. 2. Sequences *Elena* and *Elisa*. Original video frames (a), white footprints of the video atoms correlated with the correct soundtrack (b) and atoms correlated with an incorrect audio signal (c).

on such decomposition, the audio feature is extracted as described in Section 3.1. The number of basis functions used for the decomposition of audio and video signals is heuristically chosen, in order to get convenient representations. However, a distortion criteria can be easily set, in order to automatically determine the required number of atoms.

In Fig. 2, we show the results of the described procedure applied to the test sequences *Elena* and *Elisa*. Both sequences show one person speaking in front of the camera and they last about 8 seconds, i.e. they are approximately 200 frames long. In Fig. 2 (a), two frames from the original video sequences are depicted. In Fig. 2 (b), the image structures that are more correlated with the corresponding soundtrack are highlighted in white. Fig. 2 (c) illustrates the video components that are more correlated with the audio signal of a different video sequence.

The experiments show that the proposed methodology allows to clearly distinguish between correct and incorrect audio and to locate the speaker's mouth. Moreover, in all the simulations that we have run, the estimated correlation coefficients are always higher (about $10 - 20\%$) for matching audio-visual signals than for discordant audio-visual pairs. The original sequences, together with the complete set of results, are available on the author's web page [13].

5. DISCUSSION AND FUTURE WORK

In the present work, we propose a dictionary approach to audio and video representation in the context of joint audio-visual analysis. The motivation for exploring this way is mainly the observation that image sequences are typically interpreted as huge pixel intensity matrices evolving in time. The fact of considering pixel-related quantities seems to us a remarkable limiting factor, since the pixel itself is a poor source of information. Video atoms, on the other hand, represent time-evolving image structures, and their parameters intuitively describe how such structures move and change their characteristics in space and time. Furthermore, such a representation is extremely concise and easy to handle.

The results shown in this paper are obtained by making use of algorithms that have been conceived for different purposes and the features and fusion criteria we consider are, to put it mildly, rough. However, the experimental results we have obtained encourage us in pursuing in this direction. In the future, we plan to investigate more appropriate audio features and to study in details the relationship between audio and video, in order to define an accurate fusion strategy. Moreover, the stability of the video representation has to be considered, in order to improve image features tracking in complex scenes.

6. REFERENCES

[1] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Proc. of NIPS*, vol. 12, 1999.

[2] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. of NIPS*, vol. 13, 2000.

[3] T. Butz and J.-P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Signal Processing*, vol. 85, no. 5, pp. 875–902, 2005.

[4] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.

[5] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proc. of ICA*, 2003, pp. 709–714.

[6] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[7] R. Gribonval, E. Bacry, S. Mallat, P. Depalle, and X. Rodet, "Analysis of sound signals with High Resolution Matching Pursuit," in *Proc. of IEEE TFTS*, 1996, pp. 125–128.

[8] O. Divorra Escoda and P. Vandergheynst, "A Bayesian approach to video expansions on parametric over-complete 2-D dictionaries," in *Proc. of IEEE MMSP*, 2004, pp. 490–493.

[9] P. Vandergheynst and P. Frossard, "Efficient image representation by anisotropic refinement in Matching Pursuit," in *Proc. of IEEE ICASSP*, vol. 3, 2001, pp. 1757–1760.

[10] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: John Wiley & Sons, 1991.

[11] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2^{nd} ed. John Wiley & Sons, 1984.

[12] http://www.cmap.polytechnique.fr/∼bacry/LastWave/.

[13] http://lts2www.epfl.ch/∼monaci/multimodal.html.