SCHOOL OF ENGINEERING - STI
SIGNAL PROCESSING INSTITUTE
Gianluca Monaci, Òscar Divorra Escoda, Pierre Vandergheynst

ELD 241 (Bâtiment ELD)
Station 11
CH-1015 LAUSANNE

Tel: +41 21 6935615
Fax: +41 21 693 7600
e-mail: gianluca.monaci@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

ANALYSIS OF MULTIMODAL SEQUENCES USING GEOMETRIC VIDEO REPRESENTATIONS

Gianluca Monaci, Òscar Divorra Escoda and Pierre Vandergheynst

École Polytechnique Fédérale de Lausanne (EPFL)

Signal Processing Institute Technical Report

TR-ITS-2005.017

June 24, 2005

Submitted to *Signal Processing*

Analysis of Multimodal Sequences Using Geometric Video Representations

Gianluca Monaci, Òscar 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
Web page: http://lts2www.epfl.ch

Abstract

This paper presents a novel method to correlate audio and visual data generated by the same physical phenomenon, based on sparse geometric representation of video sequences. The video signal is modeled as a sum of geometric primitives evolving through time, that jointly describe the geometric and motion content of the scene. The displacement through time of relevant visual features, like the mouth of a speaker, can thus be compared with the evolution of an audio feature to assess the correspondence between acoustic and visual signals. Experiments show that the proposed approach allows to detect and track the speaker's mouth when several persons are present on the scene, in presence of distracting motion, and without prior face or mouth detection.

Index Terms

Multimodal data processing, audiovisual association, geometric video representation, sparse redundant decomposition, Matching Pursuit.

I. Introduction

Human perception of the world is essentially multimodal. We continuously combine different sensorial experiences to obtain an accurate and reliable representation of the surrounding environment, and this without any apparent effort. Concerning audio and visual data modalities, human beings have a special ability in understanding what is happening in an audiovisual scene and, given a particular sound event, in locating its origin and if it has been generated by some visible mechanical action. In other words, we are particularly efficient at assessing audiovisual synchrony. Automatic systems, on the other hand, encounter great difficulties when trying to understand relationships between audio and video signals. Providing computers with multimodal analysis capabilities is not an easy task, given the already challenging nature of "monomodal" analysis itself. The goal of this research work is that of studying and possibly understanding the relationships that exist between acoustic and visual signals, in order to detect those regions in an image sequence from which the soundtrack originates. To achieve that, we propose a new representational framework for audiovisual fusion.

The video signal is modeled as a superimposition of geometric primitives (called *atoms*) that evolve through time. The characteristics of such primitives are described by a set of parameters, whose time evolution indicates the motion of the associated visual structure. We can track in this way the evolution of relevant image features and easily define "significant events" occurring in the video channel. We will also build an audio feature based on the acoustic energy, which reflects the presence of sound in the audio channel. Based on such representation, we can define in the same way significant acoustic events. The definition of *acoustic and visual events* is the central part of the proposed method, since it allows to assess the correspondence between audio and video signals in an easy, precise but still general fashion. Audio-video signals are considered to be correlated, and thus to be generated by the same physical phenomenon, simply when a co-occurrence of acoustic and visual events is observed. The investigated approach is tested on a set of real-world sequences taken from the CUAVE database. Results demonstrate that our method allows to detect the mouth of the active speaker, when distracting motion is present, and without the need of any preprocessing step as face or mouth detector.

II. Related Work

The problem we are challenging in this work is that of correlating audio and video signals in multimedia sequences, to detect consistent audiovisual pairs that could originate from the same physical phenomenon. This problem is non trivial, since complex relationships between complex signals of different nature have to be modeled. 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 between the energy of an audio track and the value of single pixels. Since a per-pixel measure was used, the hypothesis that pixels are independent of each other conditioned on the speech signal was introduced. In [1], the mutual information is derived from the Pearson's correlation coefficient under the assumption that the joint statistics are Gaussian. 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, which is equivalent to maximum mutual information projection in the jointly Gaussian case. Nock *et al.* [3] evaluate three different algorithms for assessing audiovisual synchrony in a speaker localization context. Two of the considered methods are based on mutual information: one assumes discrete distributions and the other one considers multivariate Gaussian distributions. A third algorithm makes use of Hidden Markov Models trained on audiovisual data. Audio features are extracted from Mel-frequency cepstral coefficients, while different video features are tested: the coefficients of the discrete cosine transform and the pixel intensity change. All three algorithms require training datasets in order to build *a priori* models, like the methods proposed in [1], [2].

Recently, more general algorithms based on information theoretic features optimization have been introduced. Butz and Thiran [4] propose an approach based on Markov chains modeling audio and video signals. The audiovisual consistency is assessed by maximizing the mutual information between audio and video features, where the distributions of such features are estimated using nonparametric density estimators. For the audio, a linear combination of the power spectrum coefficients that has the biggest entropy is learnt from a dataset, while the video is represented by pixel intensity change. The audio and video joint densities are deduced by training the estimator on a set of audio-video sequences. The framework developed in [4] is used in [5], to extract optimal audio features with respect to video features. These audiovisual features are then correlated by maximizing their mutual information, in order to locate the active speaker among several candidates. A method that does not make use of any previous model training was first proposed by Fisher *et al.* [6] and has been extended in their latest work [7]. The algorithm is based on a probabilistic generation model that is used to define projection rules on maximally informative subspaces. The learnt densities are used to define the relationship between different signal modalities using a nonparametric density estimator. This approach is used to solve a conversational audiovisual correspondence problem, obtaining encouraging results. In [8], a slightly different approach is used to find, in a joint manner, an optimal modeling and fusion criteria of data. Principal Components Analysis and Independent Component Analysis are performed on audio and video features at the same time, in order to find the maximally independent audio-video subspaces, and thus extract audiovisual independent components. However, this technique is not able to deal with dynamic scenes.

In contrast to previous works, in this paper the attention is focused on modeling the observed phenomena, i.e. acoustic and visual stimuli. In particular, we propose a model of video sequences that describes concisely the content of the scene. Such a representation allows the design of intuitive and precise audiovisual fusion criteria that do not require the formulation of any complex statistical model describing the relationships between audio and video.

III. Modeling Audiovisual Phenomena

The retrieval of correlations between audio and video signals is a problem with a very high dimensionality. The goal is that of locating those spatio-temporal video regions that are interrelated with a certain audio track. In order to make this problem feasible, audiovisual data need to be modeled such that dimensionality gets reduced and only relevant signal information is used. Data modeling is thus supposed to capture the main characteristics of each signal modality that may contain information about the other modality. However, existing approaches to multimodal processing typically focus on the modeling of relationships between audio and video data, rather than on modeling the data itself.

To date, methods dealing with audiovisual fusion problems basically attempt to build complete, general and complex statistical models to capture the relationships between audio and video features. But surprisingly, the employed features are extremely simple and poorly connected with the physics of the problem, in particular for what concerns visual information. Efficient signal modeling and representation requires the use of methods able to capture particular

characteristics of each signal kind. A question that arises at this point is: why should we use a representation of video based on a basis of *deltas* (i.e. pixel wise features), if video is made of moving regions surrounded by contours with high geometrical content? Pixel-related quantities seem to us a relatively poor source of information that has a huge dimensionality, it is quite sensitive to noise and does not exploit structures in images. A very simple example can clarify this concept. If a person is moving back and forth while speaking in front of the camera, the pixel values on the mouth region change depending on the lips movements *and* on the person movement. The result is that pixel intensities evolve in an undistinguishable way.

Therefore, the idea is basically that of defining a proper model for visual signals, instead of defining a complex statistical fusion model that has, however, to find correspondences between barely meaningful features. If an accurate description of the scene is available, we can actually think of detecting consistent audiovisual pairs generated by the same phenomenon (in this case, a speaker uttering a sound), by simply observing the co-occurrence of interesting audio and video events (i.e. the presence of sound and the movement of the mouth). For particular applications, one may consider the use of adapted template based approaches for video representation (in order to model particular objects and their trajectories: lips, faces, etc...). However, for generic non-application constrained approaches, the answer seems to be that we should, indeed, use a signal model capable of exploiting video structural properties while keeping generic and flexible enough.

Such properties are introduced into the video feature extraction process, considering spatio-temporal video approximations using geometric primitives. An image sequence is decomposed in 3-D video components intended to capture geometric features (like oriented edges) and their temporal evolution. In order to represent the large variety of geometric characteristics of video features, redundant codebooks of functions have to be considered. The use of geometric video decomposition has at least two main advantages:

- Unlike the case of simple pixel-based representations, when considering image structures that evolve in time we deal with dynamic features that have a true geometrical meaning. Coming back to the example of a speaker moving back and forth, if the mouth is represented using video components that track image structures and describe their position, size and orientation variations, then we are able to better interpret what is happening in the scene.
- Geometric sparse video decompositions provide compact representations of information, allowing a considerable dimensionality reduction of the input signals. This property is particularly appealing in this context, since we have to process signals of very high dimensionality.

We will show that combining geometric-driven video features with a simple audio feature, makes it possible to define a deterministic audiovisual correspondence measure. Audiovisual pairs are considered to be correlated when we observe a temporal synchrony between "events" present in both audio and video signals, that are thus supposed to be caused by the same physical phenomenon. Events will be defined as local perturbations of an equilibrium situation, exploiting the motion information of the geometric primitives describing the scene and the energy content of the audio track. In the next Sections, we will first present the algorithm used to represent video signals using 3-D geometric functions, and then we will describe the procedures adopted to extract and correlate audio and video features.

IV. Geometric Video Representation

Natural image sequences are composed of successive projected snapshots of 3-D objects. Considering these objects to describe smooth trajectories through time, one usually assumes that image sequences are well modeled by smooth transformations of a reference frame [9]. A video sequence can thus be considered as a series of frames represented by a mixture of homogeneous regions and regular contours, where the motion is represented by smooth local deformations of those regions. Coping with regular geometric deformations necessitates the use of flexible visual primitives. In order to achieve this, we advocate the use of parametric over-complete dictionaries of basic waveforms, referred to as atoms. Local deformations are then propagated along the sequence by updating the atoms' parameter field in order to approximate the succession of frames. Assuming that an image I can be approximated with a linear combination of atoms retrieved from a redundant dictionary $\mathcal{D}_{\mathcal{V}}$ of 2-D atoms, we can write:

$$\hat{I} = \sum_{\gamma_i \in \Omega} c_{\gamma_i} g_{\gamma_i} \,, \tag{1}$$

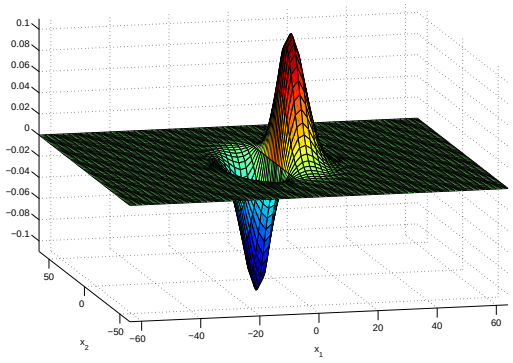


Fig. 1. The generating function $g(x_1, x_2)$ described by Eq. 5.

where i is the summation index, c_γ corresponds to the coefficient for every atom g_γ and Ω is the subset of selected atom indexes from dictionary $\mathcal{D}_\mathcal{V}$. We also require that the representation is *sparse*, i.e. the cardinality of Ω is much smaller than the dimension of the signal. The decomposition of $\hat{I}$ on an overcomplete dictionary is not unique, and several decomposition approaches have been proposed, like the method of frames [10], Matching Pursuit [11] or Basis Pursuit [12]. Here we consider Matching Pursuit (MP), an iterative greedy algorithm that selects the element of the dictionary that best matches the signal at each iteration.

Each video frame is decomposed into a low-pass part, that takes into account the smooth components of images, and a high-pass part, where most of the energy of edge discontinuities lays. The low frequency component is obtained by low-pass filtering and downsampling the images in the sequence, using the Laplacian-pyramid scheme [13]. We employ here the FIR low-pass filter proposed in [14]. The high-pass frames are obtained by subtracting the low frequency parts from the original images. These high frequency residual frames which contain the geometric structures of images, are represented using MP. At each step, MP picks up the function belonging to $\mathcal{D}_\mathcal{V}$ that best approximates the first frame of the sequence, I_1. The first step of the MP algorithm decomposes the image as

$$I_1 = \langle I_1, g_{\gamma_0} \rangle g_{\gamma_0} + R^1 I_1 \,, \tag{2}$$

where $R^1 I_1$ is the residual component after approximating I_1 in the subspace described by g_{γ_0}. The function g_{γ_0} is chosen such that the projection $|\langle I_1, g_{\gamma_0} \rangle|$ is maximal. At the next step, we simply apply the same procedure to $R^1 I_1$, which yields:

$$R^1 I_1 = \langle R^1 I_1, g_{\gamma_1} \rangle g_{\gamma_1} + R^2 I_1. \tag{3}$$

This procedure is recursively applied, and after N iterations, we can approximate I_1 as

$$\hat{I}_1 = \sum_{i=0}^{N-1} c_{\gamma_i} g_{\gamma_i} \,, \tag{4}$$

where $c_{\gamma_i} = \langle R^i I_1, g_{\gamma_i} \rangle$.

The dictionary $\mathcal{D}_\mathcal{V}$ is built by varying the parameters of a mother function, in such a way that it generates an overcomplete set of functions spanning the input image space. The choice of the generating function g is driven by the observation that it should be able to represent well edges on the 2-D plane. Thus, it should behave like a smooth scaling function in one direction and should approximate the edge along the orthogonal one. We use here an edge-detector atom with odd symmetry, that is a Gaussian along one axis and the first derivative of a Gaussian along the perpendicular one (see Fig. 1). The generating function $g(x_1, x_2)$ is thus expressed as:

$$g(x_1, x_2) = 2x_1 \cdot \mathrm{e}^{-(x_1^2 + x_2^2)}. \tag{5}$$

The codebook of functions $\mathcal{D}_\mathcal{V}$ can be defined as $\mathcal{D}_\mathcal{V} = \{g_\gamma : \gamma \in \Gamma\}$. Each atom $g_\gamma = U_\gamma g$ is built by applying a set of geometrical transformation U_γ to the mother function g. Basically, this set has to contain three transformations:

- Translations $\vec{t} = (t_1, t_2)$ all over the image plane.
- Rotations θ to locally orient the function along the edge.
- Anisotropic scaling $\vec{s} = (s_1, s_2)$ to adapt the atom to the considered image structure.

Any atom g_γ in the dictionary rotated by θ, translated by t_1 and t_2, and anisotropically scaled by s_1 and s_2 can thus be written as:

$$g_\gamma(x_1, x_2) = \frac{C}{\sqrt{s_1 s_2}} \cdot 2u \cdot \mathrm{e}^{-(u^2+v^2)}, \tag{6}$$

where C is a normalization constant and

$$u = \frac{\cos(\theta)(x_1 - t_1) + \sin(\theta)(x_2 - t_2)}{s_1}, \tag{7}$$

and

$$v = \frac{-\sin(\theta)(x_1 - t_1) + \cos(\theta)(x_2 - t_2)}{s_2}. \tag{8}$$

We consider an approach where 2D spatial primitives obtained in the expansion of a reference frame of the form of Eq. 4 are tracked from frame to frame. Given a set of images belonging to a sequence, the changes suffered from a frame I_t to I_{t+1} are modeled as the application of an operator F_t to the image I_t such that

$$\begin{aligned} I_{t+1} &= F_t(I_t), \\ I_{t+2} &= F_{t+1}(I_{t+1}) = F_{t+1}(F_t(I_t)), \\ I_{t+3} &= \dots \end{aligned} \tag{9}$$

where t is the time index.

From the model of Eq. 4 and 9, follows that

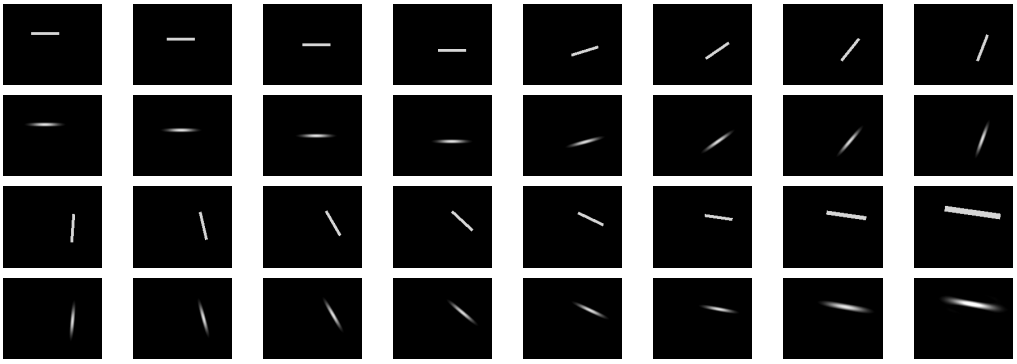
$$\hat{I}_{t+1} = F_t\left(\sum_{i=0}^{N-1} c_{\gamma_i}^t g_{\gamma_i}^t\right). \tag{10}$$

Making the hypothesis that F_t represents the set of transformations F_t^γ of all individual atoms that approximate each frame, we obtain:

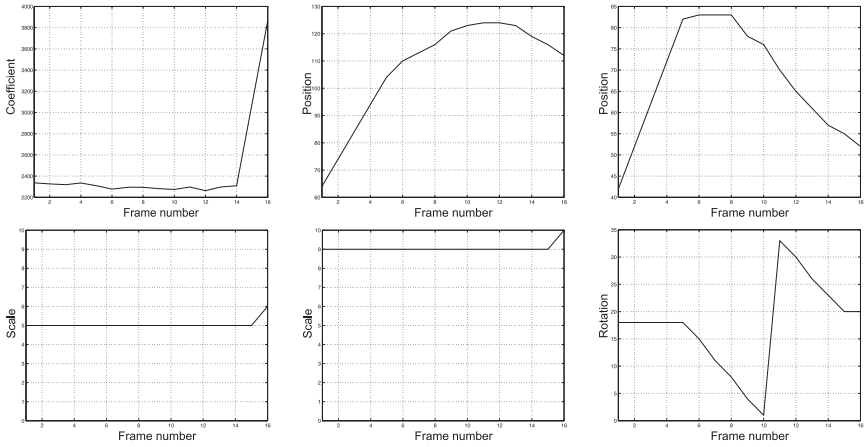
$$\hat{I}_{t+1} = \sum_{i=0}^{N-1} F_t^{\gamma_i}(c_{\gamma_i}^t g_{\gamma_i}^t). \tag{11}$$

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 parametric transformation F_t). However, at every greedy decomposition iteration some new criteria have to be considered in order to establish the relationship with the expansion of the reference frame. Only a subset of functions of the general dictionary is considered as candidate functions to represent each deformed atom. This subset is defined according to the past geometrical features of every particular atom in the previous frame, such that only a limited set of transformations (translation, scale and rotation) are possible. This imposes smoothness on the set of deformed primitives, following the assumption of smooth transformation. The formulation of the MP approach to video representation is complex and is treated in detail in [15], to which the interested readers are referred.

A cartoon example of the used approach can be seen in Fig. 2(a), where the approximation of a simple synthetic object by means of a single atom is performed. The first and third row of pictures show the original sequence and the second and fourth rows provide the approximation composed of a single geometric term. Fig. 2(b) shows the parametric representation of the sequence. We see the temporal evolution of the coefficient c_γ^t, and of the position, scale and orientation parameters. The MP decomposition of the video sequence provides a parametrization of the signal which represents the image geometrical structures *and* their evolution through time. In this way we can track the movements of relevant image features, getting an accurate description of the scene content. Besides, it is important to underline that the stream of video atoms that we consider is absolutely generic. It could be generated using different approximation techniques and it can be used to encode video sequences, as it is shown in [16].



(a) Synthetic sequence approximated by 1 atom: first and third row show the original sequence made by a simple moving object. Second and fourth row depict the different slices that form a 3-D geometric atom.



(b) Parameter evolution of the approximated object; from left to right and from up down, we find: coefficient c_γ, horizontal position t_1, vertical position t_2, short axis scale s_1, long axis scale s_2, rotation θ.

Fig. 2. Approximation of a synthetic scene by means of a 2-D time-evolving atom.

V. Audiovisual Fusion

We have now a generic representation of the video that describes synthetically how the scene is composed and how image components evolve. Using such a parametric representation, we try to follow the temporal evolution of relevant image features, like those constituting the speaker's mouth or chin. The whole set of 3-D geometric primitives used to represent the video are considered, and they are sorted by correlation with the audio. Such correspondence between acoustic and visual signals is assessed by comparing the evolution of visual structures with that of some audio track descriptors. The features considered in the following of this work are presented in Sections V-A and V-B, while the criteria used to relate them are introduced and discussed in Section V-C.

A. Audio Feature

Audio signals have a rich variety of components that human auditive system is able to perceive (Fig. 3). Correlations of the wide diversity of sounds with the also large variety of geometric configurations of the visual stimulus of a mouth are possible. Indeed, this is the main basis for *lip reading*. A positional model of lips may be assigned to each sound and transitional models between sounds can be established.

We consider here a much simpler and generic approach. As already stated, we look for synchrony between audio-video events. An interesting audio event, from our point of view, is the presence of a sound. Therefore, we need an audio feature that simply allows to assess the presence or not of an acoustic event. Finer audio features are unnecessary in this setting, but can be considered to perform more complex tasks.

Typical features used to represent audio signals are the Mel-frequency cepstral coefficients (MFCC) [17], mainly used in the speech recognition field, and employed in [2], [3], [5]. In [4] the audio feature is obtained from the spectrogram of the audio track by learning from a training dataset the linear combination of the power spectrum coefficients

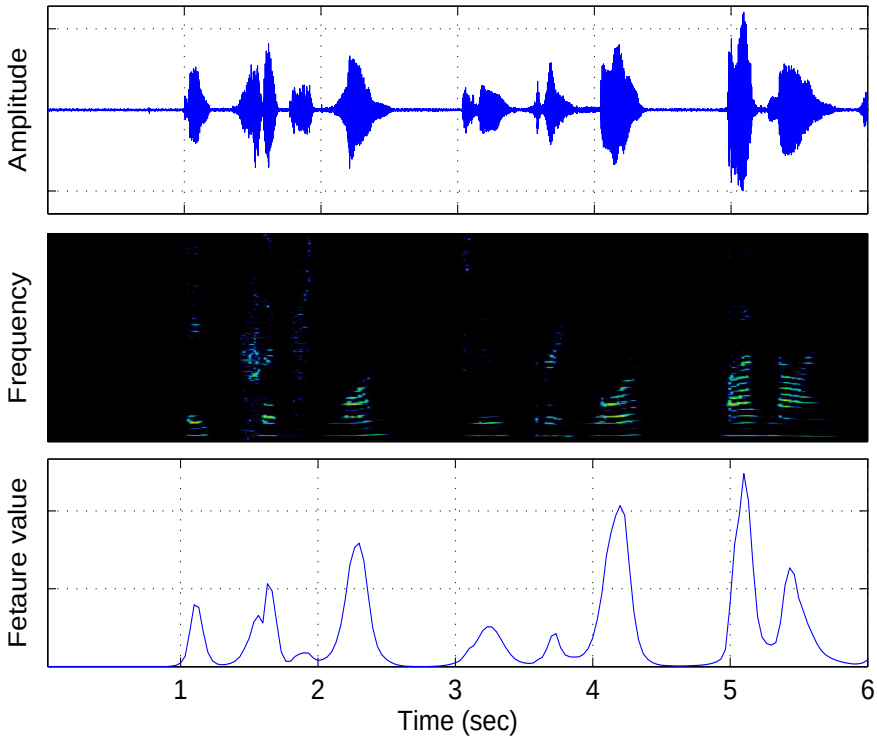


Fig. 3. Audio signal of a subject uttering eight digits in English (top), its time-frequency energy distribution $E_a(t,\omega)$ (middle), and the estimated audio feature $f_a(t)$ (bottom). The signal is decomposed using 1000 Gabor atoms. The color map of the time-frequency plane image goes from black to red, through blue, green and yellow, and the pixel intensity represents the value of the energy at each time-frequency location.

with biggest entropy. Fisher and Darrell [7] propose a similar feature that maximizes the mutual information with the video. This uses an on-line procedure that does not require a training process. In all cases, the final feature is a 1-D function that is downsampled in order to obtain the same length for audio and video features.

Here, an estimate of audio energy contained per frame is considered. To compute such an estimate, we exploit the properties of signal representations over redundant dictionaries using MP. The sparse decomposition of the audio track, in fact, performs a denoising of the signal, pointing out its most relevant structures. In the next paragraph we will briefly recall the basic steps of the MP algorithm for 1-D signals.

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

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\Big(\frac{t-u}{s}\Big) \mathrm{e}^{i\xi t}\,. \tag{12}$$

In our case, we consider a dictionary of Gabor atoms, that is, the generating function $g(t)$ is a normalized Gaussian window. The choice of a Gabor dictionary is motivated by the optimal time-frequency localization of the Gaussian core [18].

As in the case of images, after N iterations the signal $a(t)$ is represented by MP as

$$a(t) = \sum_{n=0}^{N-1} \langle R^n a, g_{\gamma_n}\rangle g_{\gamma_n} + R^N a\,, \tag{13}$$

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

An estimate of the time-frequency energy distribution of the real function $a(t)$ can be derived from its MP decomposition by summing the Wigner-Ville distributions $Wg_\gamma(t,\omega)$ of the obtained atoms [11]:

$$E_a(t,\omega) = \sum_{n=0}^{N-1} |\langle R^n a, g_{\gamma_n}\rangle|^2 \cdot Wg_{\gamma_n}(t,\omega). \tag{14}$$

If $g(t)$ is, as in that case, the Gaussian window, its Wigner-Ville distribution is

$$Wg(t,\omega) = 2 \cdot \mathrm{e}^{-2\pi(t^2+(\omega/2\pi)^2)}. \tag{15}$$

The time-frequency energy distribution $E_a(t,\omega)$ is a sum of 2-D Gaussian functions, whose positions and variances along time and frequency axes depend on the set of parameters $\gamma_n = (s_n, u_n, \xi_n)$. One of the analyzed signals and its time-frequency energy distribution are shown in Fig. 3.

2) Audio Feature Extraction: The audio representation that we obtain from the MP decomposition it is not directly exploitable to our end and has to be further processed in order to obtain a function that is comparable with the evolution of the video parameters. We require audio features composed of the same number T of samples as the MP video features. Moreover, we would like to depict the audio signal with only one time-evolving feature, in order to speed-up the computation and to simplify the problem formulation.

Our audio feature $f_a(t)$ is obtained by estimating 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 (Eq. 14):

$$f_a(t) = \sum_{n=0}^{N-1} |\langle R^n a, g_{\gamma_n}\rangle|^2 \cdot Wg_{\gamma_n}(t,0). \tag{16}$$

Note that now the Wigner-Ville distributions are projected over the time axis. The so-obtained estimate of the audio energy per time instant is downsampled, in order to get a convenient number T of time samples. Fig. 3 shows one of the analyzed audio signal with its time-frequency energy distribution and the corresponding function $f_a(t)$. In fact, our feature is similar to those described in [4], [7], with the difference that we attribute to each frequency component the same weight.

In Fig. 4, the signal of Fig. 3 and four possible audio features associated to it are depicted:

- We draw in Fig. 4(b) a feature based on the average, over a time window spanning two video frames, of the squared modulus of the audio signal.
- Fig. 4(c) shows another audio feature computed from the average over frequencies of the energy spectrogram of the signal. The spectrogram is computed as the magnitude of the windowed discrete-time Fourier transform of the signal using a sliding window. The energy distribution is given by the squared absolute value of such time-frequency function.
- Fig. 4(d) shows a third feature based on the mean over frequencies of the energy spectrogram of the audio signal after MFCC processing. In this case, the spectrogram is reconstructed after processing it using a Mel filter bank composed of 40 filters and taking the $\log_{10}$ of the output. The energy distribution is the squared absolute value of the time-frequency function.
- We draw in Fig. 4(e) the audio feature $f_a(t)$ obtained by estimating the per-frame audio energy using Eq. 16.

The four features behave similarly, and we have chosen the forth one since it exhibits a smoother and more regular profile (see Fig. 4). This is due to the sparseness and the fine time-frequency resolution of the dictionary decomposition, that allows to obtain a description that captures nicely the evolution of the audio track, filtering out most of the signal noise. Moreover, informal tests on a set of real-world sequences have confirmed our intuition, showing that slightly better audiovisual fusion results are obtained when the audio feature of Eq. 16 is used in our proposed framework.

B. Video Feature

Clearly, video features need to capture temporal variations. To date, video features used for multimodal audiovisual fusion are often based on pixel-wise intensity difference measures. In [3] and [4], the pixel intensity change measured

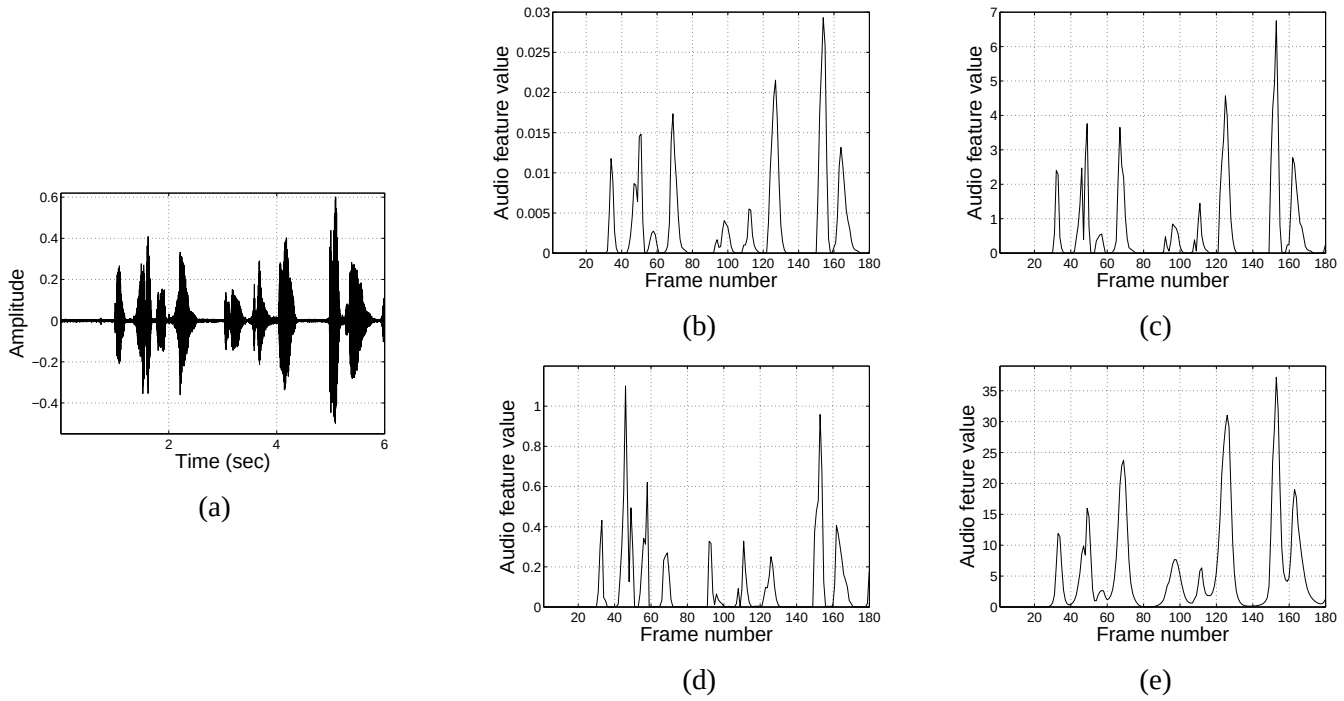


Fig. 4. The signal of Fig. 3 is drawn in (a). The average, over a time window of two video frames, of the squared modulus of the audio signal is shown in (b), the average over frequencies of the energy spectrogram of the audio signal in (c), the mean over frequencies of the energy spectrogram after MFCC processing in (d) and the per-frame audio energy estimated from the MP decomposition in (e).

in a 3×3 averaging spatial window is considered. The approach in [5] looks forward exploiting local motion information by means of optical flow measures. In any case, none of the actual approaches try to exploit the real structural nature of video signals.

We have decided thus to explore the possibilities offered by the MP video decomposition technique presented in Section IV. In this way, we hope to be able to track important geometric features over time and to effectively parameterize those transformations that represent changes in the scene. The output of the MP algorithm is a set of atom 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. 2(b) shows the atom parameters evolution as a function of time.

The video features we consider, however, are not all these 6 video parameters. The scale and orientation parameters have been discarded, since they carry few information about the mouth movements. Clearly, they can be used if needed in a more complex application, but in this context the natural choice seems that of considering a feature that takes into account the movement of image structures. Therefore, for each video atom we compute the absolute value of the displacement as

$$d = \sqrt{t_1^2 + t_2^2}, \tag{17}$$

where t_1 and t_2 are the horizontal and vertical position parameters of the atom. The quantity d is used as video feature, and indicates a sort of "activation" of the video structure that it represents. In order to be more easily compared to the audio feature, that has a smooth behavior, we convolve the video feature d with a Gaussian filter, obtaining a smooth function like the one depicted in Fig. 5 (bottom-left). Since a video sequence is represented with N time-evolving atoms, we end up with a list of N such functions composed of T time samples.

C. Fusion Criteria

Once features are available, a measure is required to determine how much these are related among them, and thus to detect those 2-D time-evolving atoms that are more correlated with the soundtrack. In the literature, different fusion criteria may be found. These are selected depending on the assumptions done to formulate the multimodal analysis problem.

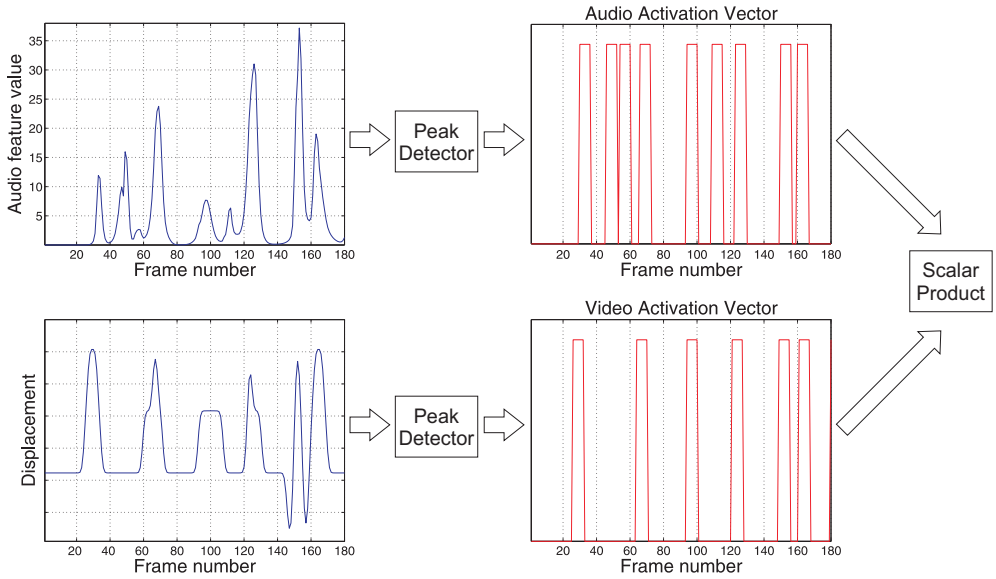


Fig. 5. Scheme of the proposed audiovisual fusion criterion. The audio energy peaks and the displacement peaks for each video atom are extracted and activation vectors are built. The Synchronization Scores between the audio activation vector and the video activation vectors are computed as the scalar product between those signals.

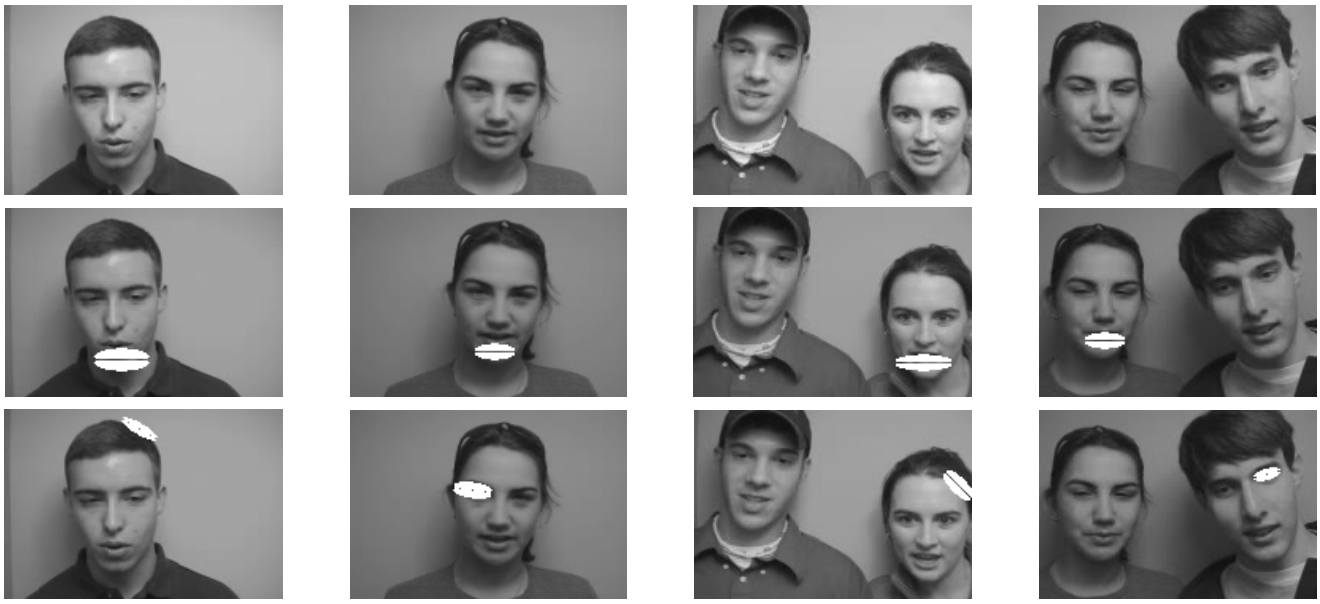
Earlier works in literature [1], [2] use *Pearson's correlation coefficient* [19] to measure the level of correspondence between acoustic and visual data. The Pearson's coefficient is a parametric measure of correlation and reflects the degree of linear relationship between two variables. The observations for both variables should be approximately (bivariate) normally distributed. We have also employed such a measure in our framework, to asses the degree of synchronization between audio and video signals [20], obtaining interesting results. However, the assumptions of linearity and Gaussianity, that are far from being satisfied in complex scenes, considerably limit the analysis power in real-world cases. Information theoretic formulations [4], [5], [7] have also proposed fusion criteria based on the use of *mutual information* [21] measure. However, when few data samples are available (as in the present case), the estimation of probability densities may pose some problems.

We propose here a very simple and powerful approach, that is derived directly from the physics of the phenomenon, and whose main steps are sketched in Fig. 5. The considered video features reflect the movement, from frame to frame, of the image structures associated with the corresponding geometric primitives. The audio feature indicates the acoustic energy content at a given time instant. Peaks in such signals suggest the presence of an event. In the video case, it can be the movement with respect to a certain equilibrium position (i.e. lips opening or closing). For the audio, a peak in the function $f_a(t)$ indicates the presence of a sound. If those audio and video peaks occur at time instants that are temporally close, we can expect that they reflect the presence of two expressions (acoustic and visual signals) of the same physical phenomenon (utterance of a sound). Therefore, peaks are extracted from the audio feature and from the N video features. Using the information about the peaks locations, we construct for each audiovisual feature function an "activation vector". Such a vector describes the presence of an event associated to the corresponding signal. It has value 1 when the feature is considered to be "active", and 0 otherwise. An activation vector $y(t)$ is built from a given feature vector $x(t)$ following the rule:

$$y(t) = \begin{cases} 1 & \text{if } x(j) \text{ is a peak, with } j - W \leq t \leq j + W \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where t is the time index. The window W models delays and uncertainty. It rarely happens, in fact, that activation peaks occur exactly at the same time instant in the acoustic and in the video feature vectors. The value of W should be chosen considering the video acquisition frame rate and the characteristics of the observed phenomenon. However, empirical observations have shown that the choice of the value is not critical. In our experiments with videos recorded at 29.97 frames per second (fps), we have obtained slightly similar results using values of W between 1 and 4.

At this point, by simply computing the scalar products between the audio and the video activation vectors built over a given observation time slot, we can assess the degree of correlation between the audio track and a 3-D atom. These



(a) (b) (c) (d)

Fig. 6. Results for *Experiment 1*: the first row shows the original video frames, the second row shows the white footprints of the video atoms correlated with the corresponding audio signal. In all cases, the speakers' mouths are correctly detected. In the third row the more correlated video atoms for a different audio source are plotted.

basic observations drive the definition of the audiovisual correlation criterion.

Definition 1: A visual primitive (3-D atom) is considered **correlated** with the audio signal over an observation time slot, if the scalar product between the corresponding activation vectors is large. This value is called **Synchronization Score**.

In here, we select the 3-D atoms that are characterized by the highest Synchronization Score with the audio, among all those that represent the video sequence.

VI. Experiments

The framework we have developed is used to locate the source of an audio signal in the video sequence. Such an application can be included in a conversational human-machine interface, in which one or more persons interact with a computer just by speaking in front of a camera, or in a smart video-conference system.

Experiments have been carried out on real-world video streams representing one or two persons speaking and moving in front of a camera. The clips used for the tests have been taken from the *individuals* and *groups* sections of the CUAVE database [22][1]. The video data was recorded at 29.97 fps and at a resolution of 480×720 pixels. The size of the clips has been then reduced to 120×176 pixels to be more easily and quickly processed. The input soundtrack was collected at 44 kHz and it was sub-sampled in order to obtain a signal at 8 kHz. The image sequence is represented using the procedure described in Section IV, obtaining a set of 2-D time-evolving atoms. The audio part is decomposed over a dictionary of Gabor atoms whose window lengths range from 512 to 16384 time samples, using the implementation of MP for 1-D signals [23] of the *LastWave* software package [24]. Based on such decompositions, the audio and video features are extracted and the activation vectors are built using a window $W = 3$, as described in Section V. The Synchronization Scores between audio and video atoms are computed and the 3-D atoms characterized by the maximum value of correlation are selected. The number of basis functions used for the decomposition of the image and audio sequences is heuristically chosen for these experiments, in order to get convenient representations. However, a distortion criteria can be easily set, to automatically determine the required number of atoms.

Two sets of experiments have been carried out. In the first series, called *Experiment 1*, we consider sequences involving only one active speaker. We have used clips consisting of one person standing in front of a camera reading digit strings, and videos involving two persons, only one of which is speaking. Each sequence lasts about 6–8 seconds.

[1]Only the luminance component of the video sequences has been considered.

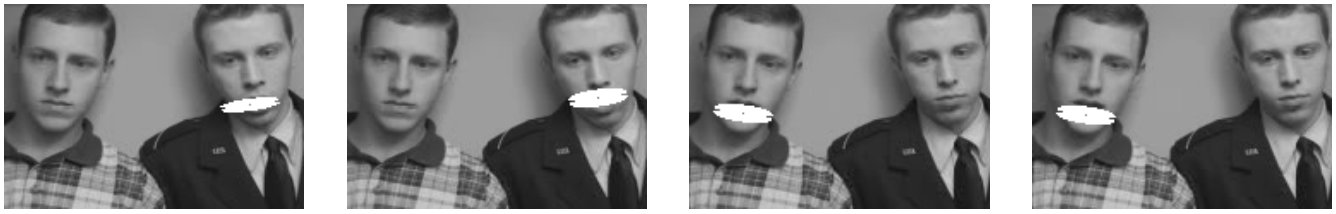


Fig. 7. Results for *Experiment 2*: four frames taken from a clip with two speakers in front of the camera taking turn in reading digit string. In the first two frames the right person is speaking, while in the last two the left one is speaking. The footprint of the most correlated atom is highlighted in white. The mouth of the correct speaker is detected.

Snapshots of some of the analyzed clips are shown in the first row of Fig. 6. We show here four non-trivial cases: speakers in sequence (a) and (b) move left and right and back and forth while uttering, the left person in clip (c) clearly mouths the text which is being pronounced by the right speaker and finally, the right subject in (d) moves significantly while the left person is speaking. In the second row of Fig. 6, the image structures that are more correlated with the corresponding soundtrack are highlighted in white. The third row of Fig. 6 illustrates the video components that are more correlated with the audio signal of a different video sequence. The audiovisual correspondence is assessed following the methodology described above and using the entire length of the sequence. Image sequences involving only one person are represented using 30 video atoms, while sequences with two subjects are represented with 50 functions. All the audio tracks are decomposed with MP using 1000 Gabor functions.

The experimental results show that the proposed methodology allows to clearly locate and track the speaker's mouth. In all the tested sequences, the algorithm chooses those visual primitives that constitute the mouth and/or chin structures of the speaker. Even when the active speaker moves, as in Fig. 6(a) and (b), or in presence of distracting motion (Fig. 6(c), (d)), the source of the sound signal is detected. On the contrary, when the video sequence is dubbed with an incongruous audio track, visual primitives which do not represent the speaker's mouth are typically detected (Fig. 6, third row). We expected such a behavior, since the proposed methodology does not simply extract moving structures, but it detects those geometric features that evolve synchronously with the audio. Finally, it is interesting to remark how video atoms adapt their orientation and shape according to the geometric characteristics of the structures they represent. Such information can be exploited in a successive stage of processing, in order to estimate the size, orientation and position of the speaker in the scene. The characteristics of the proposed approach and in particular its mouth tracking ability can be better appreciated by watching the resulting video sequences, that are available on the author's web page [25].

In *Experiment 2*, we have analyzed clips involving two active speakers, arranged as in Fig. 7. Videos show two persons taking turn in reading series of digits and last about 20 seconds. In this context, we have to introduce a sliding temporal window over which the Synchronization Scores are computed, to take into account the dynamics of the scene. A window of 2 seconds length (i.e. 60 frames) is used to detect the video atom that produces the highest Synchronization Score with the audio. The observation window is then shifted by 20 samples and the procedure repeated. Clearly, the values of window length and shift have to be chosen considering a trade-off between the response time delay of the system, and the robustness of the audiovisual association. The image sequences are represented with 50 video atoms and the audio signals are decomposed using 2000 or 3000 Gabor atoms, depending on the length of the clip. Fig. 7 shows the results of the described approach detecting the mouth of the speaker in one sequence where two persons speak in turns in front of the camera. In white are highlighted the footprints of the video atoms found to be correlated with the soundtrack. The mouth of the correct speaker is detected.

In order to quantify the accuracy of the proposed algorithm, we have manually labelled the center of the speaker's mouth in 10 sequences from the CUAVE database. The mouth of the active speaker is considered to be correctly detected if the position of the most correlated 3-D atom falls within a circle of radius r centered in the "true" mouth center. If more than one atom is selected, we estimate an atoms' centroid whose position on the image plane is given by the average of the coordinates of the single atoms. Since Synchronization Scores are computed every 20 frames, mouth labels are placed with this same frequency throughout each sequence, and performances are thus evaluated at test points distant 20 samples one from the other. In total, we have analyzed 273 test points. The values of the radius r that have been considered are 12.5 and 25 pixel. Fig. 8 shows regions of correct mouth detection for the two values of r. The green marker indicates the position of the video atom that is found to be more correlated with the audio.

(a) (b)

Fig. 8. Regions of correct mouth detection for $r = 12.5$ (a) and $r = 25$ (b). The green marker indicates the position of the most correlated video atom.

Sequence	$r = 12.5$	$r = 25$
g01	95	95
g04	86	86
g11	46	54
g12	75	82
g13	82	82
g15	83	83
g19	87	87
g20	90	93
g21	79	79
g22	87	87
Overall	81	83

TABLE I

RESULTS OF *Experiment 2* EXPRESSED IN PERCENTAGE OF CORRECT DETECTIONS.

Table I summarizes the results in term of percentage of test points at which the speaker's mouth is correctly detected. The clips are referred to with the names they have on the CUAVE dataset (g01, g04, etc...).

The values of r have been chosen in such a way that we can have a rough comparison with the results presented in [3]. Nock and colleagues propose an algorithm to detect the mouth of the active speaker founding the image region over which the mutual information between audio and video features is maximized. The method requires a training step, that is performed on the first 10 sequences of the *group* partition of CUAVE. The results are then evaluated on the remaining 12 clips, considering portions of about 20 seconds involving one single speaker. As in our algorithm, in [3] mutual information values are estimated using a sliding time window of 2 seconds that is shifted in time with steps of one second (30 samples). The goodness of the detection is assessed using the criterion that we use here, with the only difference that in [3] the speaker's mouth is considered to be correctly located if it is placed within a *square* of $T \times T$ pixel centered on the mouth center that was manually labelled. The considered values of T are 100 and 200 pixel. Thus, taking into account a downsampling factor of 4 that we have applied to the video sequences, the areas of correct mouth detection are comparable.

The set of test sequences that we consider here does not coincide completely with that used in the cited paper. However, considering those sequences analyzed in both works, we obtain considerably better performances (cfr. Table 2 in [3]). Our proposed method compares particularly favorably with Nock's one when the smaller region of correct mouth detection is considered and for challenging sequences where some distracting motion is present. For example, in clip g12, we get 75% accuracy against 46%, while for the sequence g13 we improve the 19% accuracy up to 82%. An overall mouth detection accuracy of 81% is obtained, in contrast with a 65% average accuracy measured in [3]. To be fair, we want to underline again that the considered test sets do not completely coincide, even if we have analyzed

a larger number of test points (273 in our case, 252 in the cited paper). Results clearly denote a superiority of the proposed algorithm, also considering that our correct mouth detection area is $4/\pi$ times smaller than in [3] because of the circular shape of the window. Moreover, a large fraction of errors is due to the delay introduced by the sliding observation window that causes an incorrect detection when the speaker changes. Such errors are practically imperceptible for a human observer, as can be checked observing the complete resulting sequences, that are available on the author's web page [25].

VII. Conclusions

In the present work, we propose a novel dictionary approach to audio and video representation in the context of multimodal audiovisual fusion. The motivation for exploring this way is mainly the observation that image sequences are typically interpreted as huge pixel intensities matrices evolving in time. The fact of considering pixel-related quantities seems to us a strong 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 describe concisely how such structures move and change their characteristics in space and time. This consents to handle information in an easier and faster way, and thus to develop relatively simple and intuitive, but effective, audiovisual fusion criteria.

All the work in the field use very simple representations for the signals. One evident advantage of using redundant parametric decompositions, is that we obtain a sparse representation of information, that is at the same time accurate. In our case, for example, instead of processing $120 \times 176 = 21120$ time-evolving variables (pixel intensities) to deal with the video signal, we consider only 30 or 50 variables (atoms displacements), depending on the scene's complexity. The price to pay, for the moment, is the high computational complexity of the MP algorithm, especially in what concerns the video signal. However, from our point of view this price is virtually zero, since the audio and video atoms we are using are exactly the same that the MP decoders use to reconstruct the compressed audiovisual sequence. Moreover, recent results on signal approximation show that fast algorithms for the sparse representation of signals using redundant dictionaries can be achieved [26].

We have extensively tested the proposed methodology on a large database of sequences, obtaining encouraging results. Our approach shows to be able to accurately locate the mouth of a speaker, and compares favorably with existing multimodal mouth detection algorithms. In addition, we are able in most of the cases to track visual features that have a physical meaning and indicate position, size, orientation of the image structures to which they are associated. Finally, we want again to underline that the atoms streams employed here are completely generic, could be generated by algorithms other than Matching Pursuit and can be used to encode audio and video sequences.

Acknowledgements

The authors would like to thank Lorenzo Peotta and Patricia Besson for fruitful discussions. This work is supported by the Swiss National Science Foundation through the IM.2 National Center of Competence for Research.

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] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study," in *Proc. of the 10^{th} ACM International Conference on Multimedia*, 2002.

[4] 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.

[5] P. Besson, M. Kunt, T. Butz, and J.-P. Thiran, "A multimodal approach to extract optimized audio features for speaker detection," in *Proc. of EUSIPCO*, September 2005.

[6] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. of NIPS*, vol. 13, 2000.

[7] 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, June 2004.

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

[9] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Digital Video Processing and Communications*. Prentice Hall, 2001.

[10] I. Daubechies, "Time-frequency localization operators: A geometric phase space approach,," *IEEE Transactions on Information Theory*, vol. 34, no. 4, pp. 605–612, 1988.

[11] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.
[12] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal of Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
[13] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.
[14] L. Peotta, L. Granai, and P. Vandergheynst, "Very low bit rate image coding using redundant dictionaries," in *Proc. of the SPIE, Wavelets: Applications in Signal and Image Processing X*, vol. 5207, November 2003, pp. 228–239.
[15] O. Divorra Escoda, "Toward sparse and geometry adapted video approximations," Ph.D. dissertation, EPFL, Lausanne, June 2005, [Online] Available: http://lts2www.epfl.ch/.
[16] O. Divorra Escoda and P. Vandergheynst, "A Bayesian approach to video expansions on parametric over-complete 2-D dictionaries," in *Proc. of IEEE MMSP*, September 2004, pp. 490–493.
[17] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.
[18] 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.
[19] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2^{nd} ed. John Wiley & Sons, 1984.
[20] G. Monaci, O. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," in *Proc. of IEEE ICIP*, September 2005.
[21] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: John Wiley & Sons, 1991.
[22] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus," *EURASIP JASP*, no. 11, pp. 1189–1201, 2002.
[23] E. Bacry, "LastWave software and documentation," http://www.cmap.polytechnique.fr/~bacry/LastWave/.
[24] R. Gribonval, E. Bacry, and J. Abadia, "Matching Pursuit software and documentation," http://www.cmap.polytechnique.fr/~bacry/LastWave/packages/mp/mp.html.
[25] G. Monaci, "Multimodal web page," http://lts2www.epfl.ch/~monaci/multimodal.html.
[26] P. Jost, P. Vandergheynst, and P. Frossard, "Tree-based pursuit: Algorithm and properties," Lausanne, EPFL-ITS Technical Report 2005.13, May 2005, [Online] Available: http://lts2www.epfl.ch/.