SCHOOL OF ENGINEERING - STI
SIGNAL PROCESSING INSTITUTE
*Gianluca Monaci, Pierre Vandergheynst*

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

*Tel: +41 21 693 2657*
*Fax: +41 21 693 7600*
*e-mail:* gianluca.monaci@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

# DETECTION OF SYNCHRONOUS AUDIOVISUAL EVENTS

**Gianluca Monaci and Pierre Vandergheynst**

École Polytechnique Fédérale de Lausanne (EPFL)

Signal Processing Institute Technical Report

TR-ITS-2005.36

December 6, 2005

# Detection of Synchronous Audiovisual Events

Gianluca Monaci, Pierre Vandergheynst
École Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute
CH-1015 Lausanne, Switzerland
E-mail: {gianluca.monaci,pierre.vandergheynst}@epfl.ch
Web page: http://lts2www.epfl.ch

**Abstract**

This paper presents an algorithm to correlate audio and visual data generated by the same physical phenomenon. According to psychophysical experiments, temporal synchrony strongly contributes to integrate cross-modal information in humans. Thus, we define meaningful audiovisual structures as temporally proximal audio-video *events*. Audio and video signals are represented as sparse decompositions over redundant dictionaries of functions. In this way, signals are expressed in terms of their salient structures, allowing the definition of perceptually meaningful audiovisual events. The detection of these cross-modal structures is done using a simple rule called Helmholtz principle.

Experimental results show that extracting significant synchronous audiovisual events, we can detect the existing cross-modal correlation between those signals even in presence of distracting motion and acoustic noise. These results confirm that temporal proximity between audiovisual events is a key ingredient for the integration of information across modalities and that it can be effectively exploited for the design of multi-modal analysis algorithms.

**Index Terms**

Audiovisual association, multi-modal data processing, cross-modal event localization, geometric video representation, Gestalt theory, Helmholtz principle, *a contrario* detection.

## I. Introduction

In this work we introduce and discuss a new framework for detecting events in audiovisual signals. In particular, we want to localize the source of a sound in the video sequence. Such task is quite trivial for humans, while it is particularly challenging for automatic systems. It is for this reason that we have decided to study a perceptually-driven approach to audiovisual fusion, that is based on our previous work on audio-visual modeling and fusion [1], [2], and that has been inspired by the research of Desolneux, Moisan and Morel on *Gestalt theory* and Computer Vision [3], [4], [5].

First of all, let us briefly introduce what Gestalt theory is. Starting from the first decades of past century, Gestaltists [6], [7] have tried to express all the basic laws that rule human visual perception. The basic set of such laws consists of *grouping laws*: Starting from local data, objects are formed by recursively building larger visual objects, *i.e. gestalts*, that share one or more common properties. Such properties represent specific, simple qualities of visual objects. The list of qualities according to which gestalts are built includes proximity, similarity, continuity of direction, amodal completion, closure, constant width, tendency to convexity, symmetry, common motion, past experience [7]. Clearly, such simple rules are not able, alone, to explain the human perception of the world. Thus, more complex principles governing the collaboration and the contrast between gestalt laws have also been introduced. Here, we will focus our attention on the basic set of simple grouping laws, called by Desolneux and coworkers [5] *partial gestalts*. The interested reader is referred to [7] for an exhaustive presentation of the Gestalt theory of perception.

There are two interesting facts that we want to emphasize, in order to make clear why we are interested in Gestalt theory and how is it related to our research work on cross-modal event localization.

- **Gestalt laws have been demonstrated to hold not only for visual perception, but also for other type of sensorial experiences**, like acoustic and tactile perception [7]. Moreover, several works in

psychophysics and neuroscience have also shown that Gestalt-like rules, and notably temporal proximity, contribute to integrate cross-modal information in humans [8], [9], [10], [11]. In particular, Jack and Thurlow [8] found that synchronization of visible movements with peaks of speech intensity was the main condition for considering audiovisual stimuli originated by the same generating event. Thus, we can think of designing an audiovisual event detector that exploits cross-modal information just like humans do. We will discuss more in detail in section IV how we can build a model of audiovisual phenomena that will allow us to define *meaningful audiovisual gestalts*.

- **A great effort to apply Gestalt theory to Computer Vision has been done in the last years by several researchers** [3], [4], [5], [12]. Desolneux *et al.* have shown that it exists a very simple and general principle, that they have called *Helmholtz principle*, which allows to decide whether a gestalt is reliable or not. This principle was introduced to try to describe how perception decides to group objects according to a certain quality. We will detail its formulation in section III.

To summarize, firstly we will define meaningful audiovisual gestalts. As we have just stated, one of the basic principle ruling the perception of audiovisual phenomena is the synchrony between acoustic and visual stimuli. Thus, the audiovisual structure we will consider here is the co-occurrence of an audio and a video event. Audio and video signals will be represented as sparse decompositions over redundant dictionaries of basic functions. This technique allows one to express a signal in terms of its most salient structures, making thus possible the definition of perceptually meaningful audiovisual events. Then, using the Helmholtz principle, we will detect such cross-modal gestalts.

The report is structured as follows: Section II introduces the research studies that motivated the work presented in this manuscript. Section III describes the Helmholtz principle. In section IV, we introduce the representational framework for audio and video signals and we define the meaningful audiovisual events we want to detect. Section V describes the audiovisual gestalt detection method based on the Helmholtz principle. Experimental results are reported in section VI and finally concluding remarks are given in section VII.

## II. Related Work

In this work we want to study the correlation between audio and video signals in multimedia sequences, to detect consistent audiovisual pairs that could originate from the same physical phenomenon.

Physiological and psychophysical studies have shown that audio-visual synchrony plays a fundamental role in the spatial localization of sound source. In fact, sounds appear to be produced by visual stimuli which are synchronous with acoustic signals. This effect becomes evident when the perceived spatial sound source is known to be false, as it happens when watching a show on TV or a ventriloquist's puppet. It is not thus by chance that the phenomenon of mislocating the sound source towards its apparent visual source is called in the psychophysical community *ventriloquism effect*. The phenomenon can occur in a large variety of conditions, and seems to depend strongly on the synchrony between audio and video stimuli [8], [9], [10], [11] (see also [13] for a review). Interestingly, the effect is not specific to speech, since it still appears if the lips are flipped upside-down [10] or if the mouth is replaced by synchronized light flashes [9]. What is important, is the temporal co-occurrence of audiovisual stimuli.

Hershey and Movellan [14] first used these observations to design a simple algorithm which locates sounds using audio-video synchrony. The correlation between audio and video was measured using the correlation coefficient between the energy of an audio track and the value of single pixels. Successive studies in the field [15], [16], [17], [18] focused on the statistical modeling of relationships between audio and video features, proposing more and more sophisticated, and effective, audiovisual fusion strategies. Surprisingly enough however, the audio-video features employed in these works are still extremely simple and poorly connected with the physics of the problem: We refer in particular to pixel-related features typically used for video representations. We believe that, in order to understand more in detail audio-video structures and to improve the performances of audiovisual fusion algorithms, an effort should be done to model the observed physical phenomenon. From what we have highlighted above, it seems clear that there is a predominant structure governing the processing of audiovisual signals, *i.e.* the temporal synchrony between acoustic and visual "events".

What we propose to do here is to define the sound source localization problem as a detection problem. The structure (or *gestalt*, according to the formalism introduced in the previous section) to be detected is the temporal co-occurrence of audio and video events. Clearly, the automatic definition of *meaningful* audiovisual events is non-trivial. We try to overcome this problem by using sparse representations of signals over redundant codebooks of functions intended to capture relevant signal features. In this way audio and video signals can be expressed in terms of their salient structures, and we will show that audiovisual events can be automatically and reasonably simply defined. These audiovisual gestalts will be then localized with the Helmholtz principle, which is introduced in the next section.

## III. Helmholtz Principle

The Helmholtz principle is a simple rule to decide if a partial gestalt is meaningful or not. It roughly states that an event is perceptually meaningful if it has very low probability to be observed by chance. Desolneux, Moisan and Morel formalized the Helmholtz principle in the following manner. Assume that we are observing $n$ objects $O_1, O_2, \ldots, O_n$. Assume that $k$ of them, for example $O_1, \ldots, O_k$, share a common quality. Is the presence of this common feature a pure coincidence, or is there a better explanation for it? In order to answer this question, we make the following mental experiment: We assume *a contrario* that the considered quality has been uniformly and independently distributed on all objects $O_1, \ldots, O_n$. Of course, the independence assumption is not realistic, but here we are defining an *a contrario* model which grossly represents the absence of relevant events. Then we (mentally) assume that the observed objects are distributed according to this random uniform process. Finally, we ask the question: Is the observed set of points probable or not? The Helmholtz principle states that if the expectation of the observed configuration $O_1, \ldots, O_k$ is very small, then we are observing a meaningful event, a gestalt.

The power of the Helmholtz principle resides in the fact that, conversely to classical Bayesian methods, it does not require a precise modelization of the observed phenomenon. In fact, here we coarsely model a general statistical background which represents the absence of significant events. An event is considered to be relevant if it has, according to this generic model, a very low probability. In this case, we suppose that such a particular event has a better explanation than chance alone, it is a meaningful gestalt. It is important to underline that the configurations to be detected have to be specified before the observation. Moreover, these events have to be defined so that they correspond qualitatively to some perceptually meaningful structures. We will see in the next section how this can be achieved in the case of audiovisual scenes.

## IV. Audiovisual Gestalts

As already stated, the audiovisual gestalt we want to detect is the co-occurrence of an acoustic and a visual event. Such synchronization of events is the main manifestation of a physical phenomenon (utterance of a sound by a speaker for example), whose effects are recorded over different channels (audio and video in this case). As underlined at the end of the previous section, the audiovisual configuration to detect has to be defined in such a manner that it depicts a perceptually meaningful structure. We observe here that a visual signal is mainly made of moving regions surrounded by contours with high geometrical content. Pixel-related quantities seem thus a relatively poor source of information that moreover has a huge dimensionality and does not exploit structures in images.

Therefore, the idea is that of 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. Note that representing the video signal as a set of edge-like features that are tracked trough time, we try to define meaningful video structures that obey Gestalt principles. In particular, sets of individual pixels are grouped together and represented with a 3-D moving edge according to the rules of proximity, similarity and common motion, which are three of the basic Gestalt grouping laws postulated by Kanizsa [7] (see section I).

The video representation algorithm has been developed by Divorra [19], and it has been already adopted in [1], [2], giving encouraging results in the context of multimodal sequence analysis. The use of geometric video decomposition has at least two main advantages. Firstly, when considering image structures that evolve in time we deal with dynamic features that have a true geometrical meaning. Secondly, 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.

In the next two sections, we will briefly describe the technique used to represent the audio signal and the video representation algorithm of Divorra, letting the interested reader refer to the above cited papers [1], [2], [19]. Finally, based on such representations, in section IV-C we will define meaningful audiovisual events.

### *A. Audio Representation*

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. Here, we consider an estimate of audio energy contained per frame. To compute such an estimate, we exploit the properties of signal representations over redundant dictionaries using Matching Pursuits [20] (MP). The sparse decomposition of the audio track, in fact, performs a denoising of the signal, pointing out its most relevant structures.

The audio signal $a(t)$ is decomposed using the MP algorithm of Mallat and Zhang 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, translating in time and modulating in frequency a generating function $g(t) \in L^2(R)$. 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 [21].

The approximation of $a(t)$ using basic functions taken from the codebook $\mathcal{D}_{\mathcal{A}}$ can be expressed as:

$$a(t) \approx \sum_{\omega_i \in \Omega} c_{\omega_i} g_{\omega_i}(t), \tag{1}$$

where $c_{\omega_i}$ are the coefficients and $\Omega$ is the set of atom indexes picked to approximate the signal.

An estimate of the time-frequency energy distribution of the function $a(t)$ can be derived straightforwardly from its MP decomposition [20]. From this energy distribution of the audio signal, we can derive an audio feature $f_a(t)$ that estimates the average acoustic energy present at each time instant, as we have shown in [2]. Fig. 1 shows one of the analyzed audio signal with its time-frequency energy distribution and the corresponding function $f_a(t)$.

### *B. Video Representation*

The image sequence is represented using the algorithm proposed by Divorra [19]. 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.

$I(\vec{x})$ can be approximated with a linear combination of atoms $G_\gamma(\vec{x})$ retrieved from a redundant dictionary $\mathcal{D}_{\mathcal{V}}$ of 2-D atoms, we can write:

$$I(\vec{x}) \approx \sum_{\gamma_j \in \Gamma} c_{\gamma_j} G_{\gamma_j}(\vec{x})\,, \tag{2}$$

where $j$ is the summation index, $c_\gamma$ corresponds to the coefficient for every atom $G_\gamma$ and $\Gamma$ is the subset of selected atom indexes from dictionary $\mathcal{D}_{\mathcal{V}}$. The codebook $\mathcal{D}_{\mathcal{V}}$ is built by applying a set of geometric transformations to a mother function $G$, in such a way that it generates an overcomplete set of functions spanning the input image space. The considered transformations are anisotropic scaling $s_1$ and $s_2$, translations $t_1$ and $t_2$ over the 2-D plane and rotation $\theta$. The generating function should be able to represent well edges on the

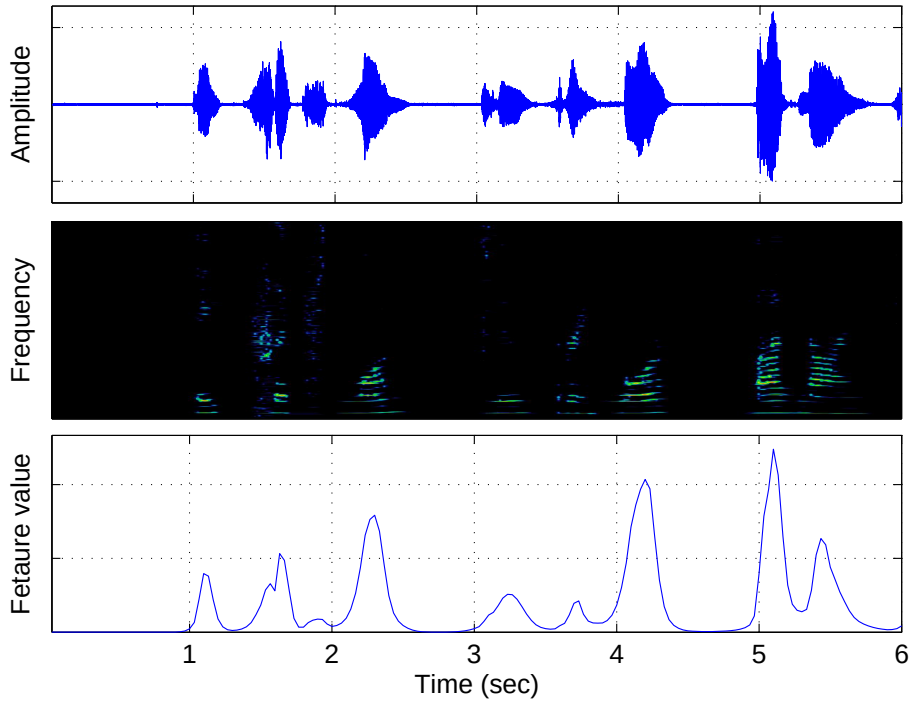



Fig. 1. Audio signal of a subject uttering eight digits in English (top), its time-frequency energy distribution, and the estimated audio feature $f_a(t)$ (bottom). The signal is decomposed using 1000 Gabor atoms. The color map of the time-frequency plane image goes from black to red, through blue, green and yellow, and the pixel intensity represents the value of the energy at each time-frequency location.

2-D plane and 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.

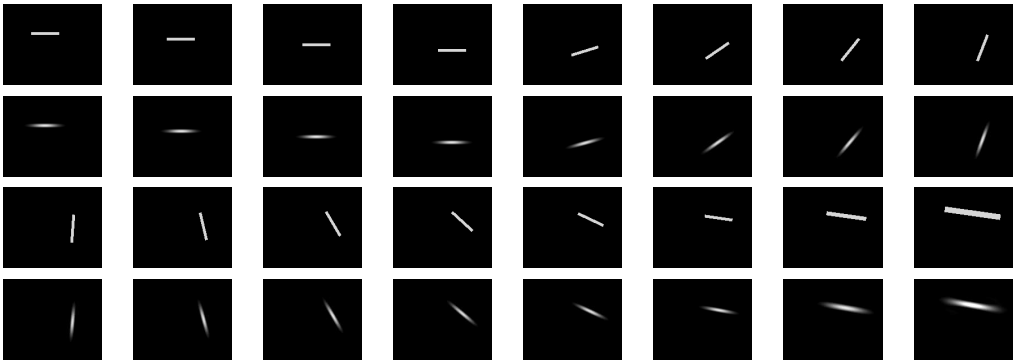
The changes suffered from a frame $I_t$ to $I_{t+1}$ are modeled as the application of an operator $F_t$ to the image $I_t$ such that $I_{t+1} = F_t(I_t)$ and

$$I_{t+1}(\vec{x}) = \sum_{\gamma_j \in \Gamma} F_t^{\gamma_j} \cdot (c_{\gamma_j}^t G_{\gamma_j}^t(\vec{x}))\,, \tag{3}$$

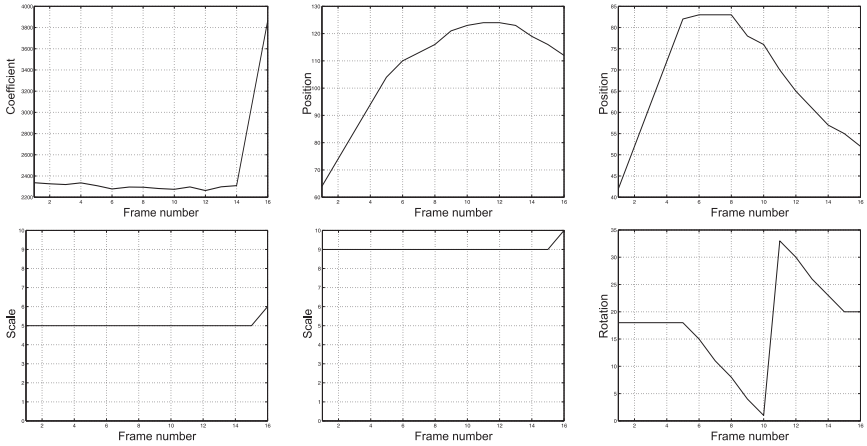
where $F_t$ represents the set of transformations $F_t^\gamma$ of all atoms that approximate each frame. A MP-like approach similar to that used for the first frame is applied to retrieve the new set of $G_\gamma^{t+1}(\vec{x})$ (and the associated transformation $F_t$). At every greedy decomposition 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 [19], to which the interested reader is 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 $c_\gamma^t$, 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.

### C. Meaningful Audiovisual Events

The audiovisual structure we want to detect is the synchrony between movements in the video and sound peaks in the audio signal. The audio feature $f_a(t)$, depicted in Fig. 3 (b), basically estimates the average



(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 bar. From left to right and from up down, we find: Coefficient $c_\gamma$, horizontal position $t_1$, vertical position $t_2$, short axis scale $s_1$, long axis scale $s_2$, rotation $\theta$.

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

energy present in the audio signal $a(t)$. The output of the MP video algorithm, instead, 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 (see Fig. 2(b)). From the position parameters, we can compute the displacement of each video atom and thus extract exactly the information we desire, that is the movement of important visual structures. Therefore, for each video atom we compute the absolute value of the displacement as

$$d = \sqrt{t_1^2 + t_2^2}, \tag{4}$$

where $t_1$ and $t_2$ are the horizontal and vertical position parameters of the atom. 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 kernel, obtaining a smooth function like the one depicted in Fig. 3 (c).

At this point, we have one audio feature and $N$ video features that describe the movement of relevant visual features, where $N$ is the number of atoms used to represent the video sequence. Each of these variables have the same number of samples $T$, since we downsample $f_t(a)$ that has a higher temporal resolution.

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 are in the presence of a *gestalt* that reflects two expressions (acoustic and visual

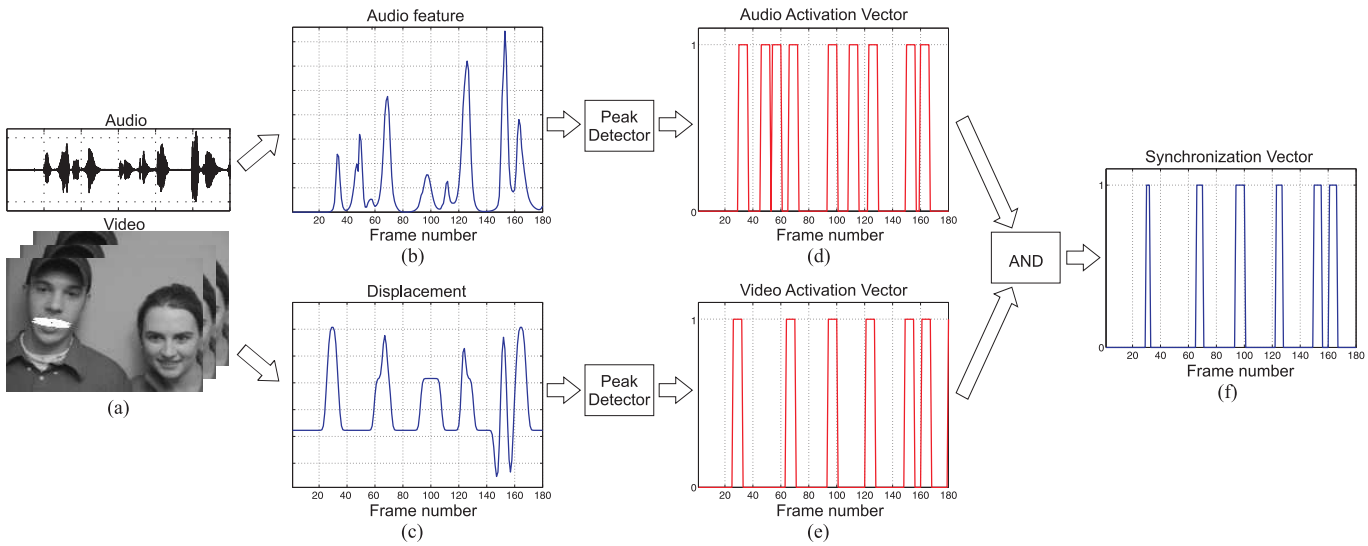



Fig. 3. Scheme of the proposed audiovisual fusion criterion. Starting from the original audiovisual sequence (a), we compute the audio feature $f_a(t)$, shown in (b), and the displacement feature associated to a video atom placed over the speaker's mouth, depicted in (c). The evolution of the two features exhibit a remarkable synchrony. From these signals we extract the audio energy peaks and the displacement peaks and the activation vectors $y_a(t)$ and $y_v(t)$ are built (d–e). The synchronization vector $s(t)$ is constructed by computing the logical AND between the two activation vectors (f).

signals) of the same physical phenomenon (production of a sound). Thus, for a given feature vector $x(t)$ we build an *activation vector* $y(t)$ which is based on the information about the peaks locations. First, we detect the peaks in the audio feature and in each of the $N$ video features, obtaining vectors which equal 1 where peaks occur and 0 otherwise. Then, such vectors are filtered with a rectangular window of size $W$. The filter models delays and uncertainty, since it rarely happens that activation peaks occur exactly at the same time instant in both acoustic and video feature vectors. An activation vector describes the presence of an event associated to the corresponding signal. It has value 1 when the feature is "active", and 0 otherwise.

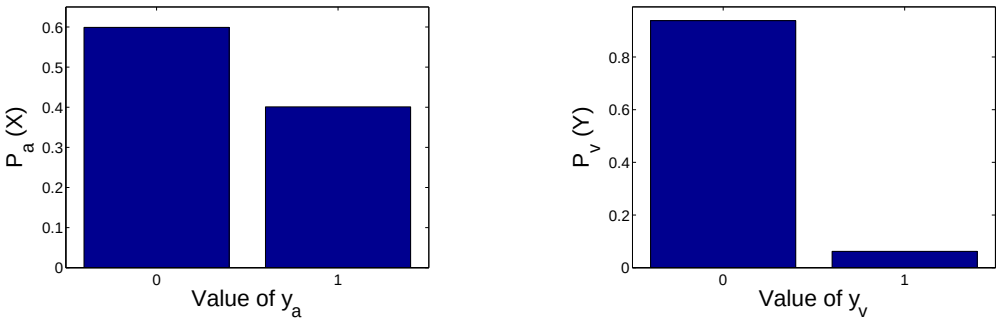
We end up with one activation vector for the audio, $y_a(t)$, and $N$ activation vectors $y_v^i(t)$, one for each video atom. By simply computing a logical AND between $y_a(t)$ and all the video activation vectors constructed over a given observation time slot, we build $N$ vectors, that we call *synchronization vectors* $s_i(t)$. The vectors $s_i(t)$ keep value 1 at those time instants at which both audio and the considered video atom are active and 0 otherwise. Thus, the number of 1 present in the vector indicates the degree of synchronization between the audiovisual pair. Fig. 3 summarizes the construction of one synchronization vector $s_i(t)$.

## V. Detection of Audiovisual Meaningful Events

Once synchronization vectors are available, we need a method to select those audiovisual structures which form *meaningful* audio-video pairs. We would like to do that in an automatic way, and tuning as less parameters as possible. In the next sections we will show how we can build a multi-modal event detector based on the Helmholtz grouping law presented in section III. The parameters of the algorithm reduce to just one, from which the detection accuracy weakly depends.

### A. An Audiovisual Event Detector Based on the Helmholtz Principle

At this step of the reasoning, for each video atom we have built a synchronization vector $s_i(t)$. Now, suppose that we observe a synchronization vector of length $n$ (*i.e.* it has been built over a temporal window of $n$ samples), and let the number of 1 in such vector be equal to $k$. We can ask ourselves: Is the number $k$ big enough, so that we can consider the corresponding video atom correlated with the audio signal? Or the co-occurrence of audio and video events is due only to chance? We can try to answer to these questions by applying the Helmholtz principle.




(a) (b)

Fig. 4. Empirical distributions of $y_a(t)$ (a) and of one feature $y_v^i(t)$, with $i = 1, \ldots, N$ (b). The normalized frequency histograms depicted here are computed for the test sequence Movie 2.

We first have to define the background *a contrario* model, which corresponds to the absence of correlated audiovisual events. In this case it is sound to consider that the observations $y_a(t)$ and $y_v^i(t)$ are independently, identically distributed random variables. Since the general form of their distributions are unknown (anyway, it is not reasonable to assume that a single distribution could account for all audiovisual sequences), the empirical distributions are considered. Integrating the empirical distribution functions (frequency histograms) yields the functions $P_a(X)$ and $P_v(Y)$, where $X$ and $Y$ are random variables distributed according to the empirical distributions of the observed values $y_a(t)$ and $y_v^i(t)$ (with $i = 1, \ldots, N$) respectively. Fig. 4 shows the empirical distributions of $y_a(t)$ and $y_v^i(t)$ computed for one of the test sequences.

Let A be a 3-D atom with corresponding synchronization vector $s_{\mathsf{A}}$ of length $n$, and let $k$ be the number of points at which $s_{\mathsf{A}}$ assumes value 1. Let us define the event $E =$ *"At least $k$ points of a vector of size $n$ have a value of $s_{\mathsf{A}}$ equal to 1"*. Note that the probability $P(s_{\mathsf{A}} = 1)$, given the independence hypothesis, can be written as

$$P(s_{\mathsf{A}} = 1) = P_a(X = 1) \cdot P_v(Y = 1). \tag{5}$$

Thus, according to the background model, the probability of the event $E$, $P(E)$, is

$$P(E) = \mathcal{B}(k, n, P(s_{\mathsf{A}} = 1)) = \mathcal{B}(k, n, P_a(X = 1) \cdot P_v(Y = 1)) \tag{6}$$

where $\mathcal{B}(k, n, p)$ is the tail of a binomial distribution:

$$\mathcal{B}(k, n, p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}. \tag{7}$$

According to these notions, we can now define an $\varepsilon$-*meaningful* video atom. Let us stress that in this context, the meaningfulness of a 3-D atom is referred to its correlation with the audio signal.

*Definition 1:* For a given atom A with corresponding synchronization vector $s_{\mathsf{A}}$ of size $n$ and containing $k$ matching points (*i.e.* $k$ values equal to 1), we define the "number of false alarms" (*NFA*) as:

$$NFA(\mathsf{A}) = N \cdot \mathcal{B}(k, n, P(s_{\mathsf{A}} = 1)), \tag{8}$$

where $N$ is the number of tests.

A 3-D atom A is said to be $\varepsilon$-meaningful if $NFA(\mathsf{A}) \leq \varepsilon$.

It is easy to demonstrate that the expected number of $\varepsilon$-meaningful 3-D atoms in a sequence, according to the *a contrario* model, is less then $\varepsilon$ [3], [12]. Moreover, it also possible to show that the number $k$ of matching points in a synchronization vector that are required to be significative depends on the logarithm of $\varepsilon$ and $N$ [3], [12]. This means that the detection results are robust to variations of those values.


Fig. 5. Movie 1: Audio signal (left), sample raw frame (center) and corresponding dynamic pixels (right). Gray-levels on the right picture represent the absolute value of the difference image between two successive frames. Black pixels thus represent static regions.

### B. Setting of the Parameter $\varepsilon$

The value of $\varepsilon$ controls the number of false detections. Setting $\varepsilon$ equal to 1, as done in [12], means that the expected number of false detection in a sequence distributed according to the background model is less than 1. However, the hypothesis of independence, especially for what concerns the video representation, is far from being realistic since the video atomic decomposition exploits the correlation between neighboring atoms (see [2], [19] for details). Because of this, several video atoms exhibit $NFA$ smaller then $\varepsilon = 1$, even without being correlated with the audio. One solution is that of considering a value of $\varepsilon$ that is smaller then 1, as it is done in [3] where $\varepsilon = 1/10$.

However, better results can be achieved by exploiting some additional knowledge about the scene. Here, we are implicitly assuming that a single audiovisual source is observed at each time instant. Thus, the solution we want to find should be well localized in the image plane. Following this reasoning, we can test multiple values of $\varepsilon$ (smaller than 1), keeping the solution which is more localized in space. By doing that, we basically do not fix any detection threshold. Instead, we browse a set of interesting solutions and we chose the most suitable one.

In practice, what we will do is to consider a set of $\varepsilon_i$ uniformly spaced in a logarithmic scale between $\varepsilon_{MIN}$ and 1. For each value $\varepsilon_i$, we obtain a set of video atoms $G_i$ for which $NFA(\mathsf{A}) \leq \varepsilon_i$, with $\mathsf{A} \in G_i$. For each group $G_i$, the variances along the horizontal ($\text{var}_x$) and vertical positions ($\text{var}_y$) are computed and the maximum value $V_{G_i} = \max\{\text{var}_x(G_i), \text{var}_y(G_i)\}$ is kept. Clearly, a set of video atoms can be composed of only one function: In that case the variance $V_{G_i}$ is equal to zero. If a group is empty, its variance is set to a very high value (ideally infinite). This is done to avoid the algorithm to search for a very small threshold $\varepsilon_i$ for which the corresponding group $G_i$ is empty and has thus zero variance. Our considered solution $G^*$ is the set of atoms which exhibits the smallest variance $V_{G^*}$.

## VI. Experiments

We show here how the proposed framework is used to locate the source of an audio signal in real-world video sequences.

The first video clip, Movie 1, shows a hand playing the piano while a toy car moves. It was recorded at 25 frames/sec with a resolution of 144 $\times$ 180 pixels. Movie 2, instead, shows two persons taking turn in reading series of digits. The video data was sampled at 29.97 frames/sec and it has a resolution of 120 $\times$ 176 pixels. Both soundtracks were collected at 44 kHz and sub-sampled to 8 kHz. For both sequences, only the luminance component is considered. The original soundtrack and a sample raw frame of Movie 1 are depicted in Fig. 5.

The image sequences are represented with 50 time-evolving atoms, while the audio track is decomposed using 1000 Gabor atoms using the implementation of MP for 1-D signals of the *LastWave* software package [22]. Based on such decompositions, the audio and video features are extracted and the activation vectors are built using a window of size $W = 7$. The synchronization vectors are computed

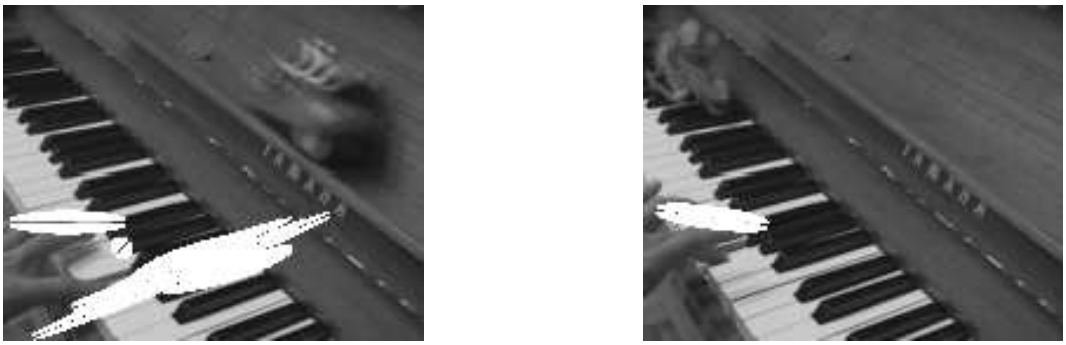


Fig. 6. Results of the algorithm run on Movie 1. Correlated atoms, highlighted in white, represent the player's fingers and the piano keys. The moving toy car, instead, is not detected.

and the set of meaningful 3-D atoms $G^*$ is selected using $\varepsilon_{MIN} = 10^{-6}$ and the set of thresholds $\varepsilon_i = \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The number of basis functions used to represent the image and audio sequences is heuristically chosen, in order to get convenient representations. However, a distortion criteria can be easily set, to automatically determine the required number of atoms.

For the analysis of the sequences, we use a sliding window of 60 frames over which the synchronization vectors are computed, in order to take into account the dynamics of the scene. At each step the observation window is shifted by 20 samples and the procedure iterated. The values of window length and shift have been chosen considering a trade-off between the response time delay of the system, and the robustness of the association. However, the algorithm is basically parameter-free since all the values that have to be set are fixed for all the experiments. Moreover, from informal tests, the choice of none of the parameters results to be critical.

Fig. 6 shows resulting sample frames of the algorithm run on Movie 1. In white, the footprints of the video atoms correlated with the soundtrack are highlighted. The player's fingers and the piano keys are detected as sound source. In the left picture several 3-D atoms are extracted, since different keys are touched in rapid succession, while in the right image one player's finger is detected. The moving toy car introduces a considerable distracting motion (see Fig. 5) and a non-negligible acoustic noise. However, it is filtered out by the cross-modal localization algorithm.

Fig. 7 shows similar results for Movie 2. In the first two sample frames the left person is speaking, while in the last two the right one is. The sequence is non-trivial, since in the second part of the movie the left person mouths the digits which are being uttered by the right speaker. However, the algorithm is able to correctly localize the mouth and the chin of the current speaker. It is interesting to remark how video atoms adapt their orientation and shape according to the geometric characteristics of the structures they represent.

Fig. 7. Results for Movie 2: In the first two sample frames the left person is speaking, while in the last two the right one is. The most correlated 3-D atoms are highlighted in white. The mouth and the chin of the correct speaker are detected.

## VII. Conclusions

In this paper we present a novel algorithm for the cross-modal fusion of audiovisual signals. Multi-modal signals are decomposed over redundant dictionaries of atoms, obtaining concise representations that

moreover describe the structural properties of those signals. This allows to define meaningful audio-video events (*gestalts*) that can be detected using a simple rule, the Helmholtz principle.

The proposed audiovisual events detection method features several interesting properties:

- **The algorithm exploits the inherent physical structures of the observed phenomenon**. This allows the design of intuitive but effective audiovisual fusion criteria and demonstrates that temporal proximity between audiovisual events is a key ingredient for cross-modal integration of information. In addition to its simplicity, the proposed method also exhibits robustness to significant audio-video distractors.
- **The algorithm naturally deals with dynamic scenes**.
- **There is no parameter to tune**. All the parameters are fixed and from informal tests, the algorithm results robust to significant variations of their values.
- **Visual information is described in a very concise fashion**. For example, instead of processing $144 \times 180 = 25960$ time-evolving variables (pixel intensities), we consider only 50 variables (atoms displacements).
- **The atoms streams employed here are completely general**, could be generated by algorithms other than MP and can be used to encode the audio and video sequences.
- **The description of the scene is extremely rich**. The audio and video atomic decompositions bring a large amount of information that can be exploited at different processing levels. If needed, for example, the information about scale and orientation of the video atoms can be exploited.

The price to pay, for the moment, is the high computational complexity of the MP algorithm. However, recent results on sparse signal approximation show that fast methods for the representation of signals over redundant codebooks of functions can be achieved [23].

Possible extensions of this work include the use of stereo sound to improve the spatial localization capabilities of our approach and possibly to extend it to the multiple sources case. Moreover, we are investigating the possibility of applying our technique to other types of multimodal signals, like climatologic data or data from robot sensors (*e.g.* terrain images and inertial sensors).

### *Acknowledgements*

The authors are grateful to Òscar Divorra Escoda for fruitful discussions. We also thank Frédéric Cao for having introduced us to Gestalt theory and for having read this manuscript. This work is supported by the Swiss National Science Foundation through the IM.2 National Center of Competence for Research.

## REFERENCES

[1] G. Monaci, Ò. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal signals using redundant representations," in *Proc. of IEEE ICIP*, September 2005.

[2] G. Monaci, Ò. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Processing, in press*, 2006, [Online] Available: `http://lts2www.epfl.ch/`.

[3] A. Desolneux, L. Moisan, and J.-M. Morel, "Meaningful alignments," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 7–23, 2000.

[4] A. Desolneux, L. Moisan, and J.-M. Morel, "Edge detection by Helmholtz principle," *Journal of Mathematical Imaging and Vision*, vol. 14, no. 3, pp. 271–284, 2001.

[5] A. Desolneux, L. Moisan, and J.-M. Morel, "A grouping principle and four applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 4, pp. 508–513, 2003.

[6] M. Wertheimer, "Untersuchungen zur lehre der gestalt, II," *Psychologische Forschung*, vol. 4, pp. 301–350, 1923, Translation published as "Laws of Organization in Perceptual Forms", in: W. Ellis, *A Source Book of Gestalt Psychology*, pp. 71-88, Routledge and Kegan Paul, London, 1938.

[7] G. Kanizsa, *Grammatica del vedere. Saggi su percezione e gestalt*, Il Mulino, Bologna, 1980.

[8] C. E. Jack and W. R. Thurlow, "Effects of degree of visual association and angle of displacement on the "ventriloquism" effect," *Perceptual and Motor Skills*, vol. 38, pp. 976–979, 1973.

[9] M. Radeau and P. Bertelson, "Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations," *Perception and Psychophysics*, , no. 22, pp. 137–146, 1977.

[10] P. Bertelson, J. Vroomen, G. Wiegeraad, and B. deGelder, "Exploring the relation between McGurk interference and ventriloquism," in *Proc. of International Conference on Spoken Language Processing*, 1994, vol. 2, pp. 559–562.

[11] M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo, "Unifying multisensory signals across time and space," *Experimental Brain Research*, vol. 158, no. 2, pp. 252–258, 2004.

[12] F. Cao, "Application of the Gestalt principles to the detection of good continuations and corners in image level lines," *Computing and Visualization in Science*, vol. 7, pp. 3–13, 2004.
[13] P. Bertelson, "Ventriloquism: A case of cross-modal perceptual grouping," in *Cognitive Contributions to the Perception of Spatial and Temporal Events*, G. Aschersleben, T. Bachmann, and J. Müsseler, Eds., pp. 347–362. Elsevier, 1999.
[14] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Proc. of NIPS*, 1999, vol. 12.
[15] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. of NIPS*, 2000, vol. 13.
[16] 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.
[17] 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.
[18] 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.
[19] Ò. Divorra Escoda, *Toward Sparse and Geometry Adapted Video Approximations*, Ph.D. thesis, EPFL, Lausanne, June 2005, [Online] Available: `http://lts2www.epfl.ch/`.
[20] S. Mallat and Z. Zhang, "Matching Pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.
[21] R. Gribonval, E. Bacry, S. Mallat, Ph. Depalle, and X. Rodet, "Analysis of sound signals with High Resolution Matching Pursuit," in *Proc. of IEEE TFTS*, 1996, pp. 125–128.
[22] R. Gribonval, E. Bacry, and J. Abadia, "Matching Pursuit software and documentation," `http://www.cmap.polytechnique.fr/~bacry/LastWave/packages/mp/mp.html`.
[23] P. Jost, P. Vandergheynst, and P. Frossard, "Tree-based pursuit: Algorithm and properties," EPFL-ITS Technical Report 2005.13, Lausanne, May 2005, [Online] Available: `http://lts2www.epfl.ch/`.