Audiovisual Gestalts

Gianluca Monaci, Pierre Vandergheynst
Signal Processing Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{gianluca.monaci,pierre.vandergheynst}@epfl.ch - 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, it is possible to define 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.

1. Introduction

Humans continuously combine audio and video stimuli to enhance their perception of the world. In fact it has been shown that sounds appear to be produced by visual stimuli which are synchronous with acoustic signals. The phenomenon can occur in a large variety of conditions, and it seems to depend strongly on the synchrony between audio and video stimuli [6, 19].

These observations motivated Hershey and Movellan [8] to design a simple algorithm to locate 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 [17, 14, 18, 7, 11] focused on the statistical modeling of relationships between audio and video features, proposing audiovisual fusion strategies based on Canonical Correlation Analysis [17, 11], Independent Subspace Projections [18] and Mutual Information maximization [14, 7]. Surprisingly enough, the audio-video features employed in these works are still extremely simple and barely connected with the physics of the problem: we refer in particular to pixel-related features typically used for video representations. This makes it difficult to deal with dynamic scenes, since the variables that are observed (pixel values or related quantities) are static. Moreover, pixel-related values have low semantic content, which makes it impractical to extract and manipulate correlated audiovisual structures.

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. In this work we introduce a new framework for detecting meaningful events in audiovisual signals. In particular, we want to localize and extract the source of a sound in a video sequence. Methods exist that perform this task using multi-microphone systems and stereo triangulation to estimate the spatial location of sounds [1, 16]. Instead, we want to achieve that using an image sequence and one microphone, exploiting thus the correlation between audio and video signals. We propose here a perception-inspired approach to audiovisual fusion that is based on previous work on multi-modal analysis by Monaci *et al.* [13], and which is inspired by the research of Desolneux *et al.* on *Gestalt theory* and Computer Vision [3, 4].

Starting from the first decades of past century, Gestaltists [10] have tried to express the basic laws ruling 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. The list of qualities according to which gestalts are built includes proximity, similarity, continuity of direction, common motion, closure, symmetry, past experience [10].

There are two interesting facts that we want to underline, in order to clarify why we are interested in Gestalt theory and how this is related to cross-modal event localization.

- As we have recalled above, Gestalt-like rules and notably temporal proximity, strongly contribute to the integration of cross-modal information [6, 19]. 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 Sec. 3 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 was done in the last years by several researchers [2, 3, 4]. Desolneux *et al*. [3] introduced a very simple and general rule, 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 groups objects according to a certain quality. We will detail its formulation in the next section.

In this paper we improve the method in [13] by formalizing the audiovisual fusion task as a gestalt detection problem. The resulting algorithm is elegant and basically free of user defined parameters, and it allows to intuitively extract and handle correlated audiovisual components with high semantic meaning. Audiovisual gestalts are defined as co-occurrences of audio and video events. Audio and video signals are sparsely decomposed over redundant dictionaries of functions. In particular, video sequences are expressed as sums of time-evolving visual structures, allowing to naturally handle dynamic scenes. Using sparse decompositions, a signal can be represented 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 performances of the proposed approach are demonstrated in real-world sequences and they are compared with those of existing sound localization algorithms.

2. 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 *et al*. [3] formalized this principle in the following manner. Assume that we are observing n objects $O_1, \dots, O_n$. Assume that k of them, $O_1, \dots, O_k$, share a common quality. Is the presence of this common feature a coincidence, or is there a better explanation for it? To answer this question, we do this mental experiment: we assume *a contrario* that the considered quality was uniformly and independently distributed on all objects $O_1, \dots, O_n$. Clearly, 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 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 $O_1, \dots, O_k$ is small, then we are observing a meaningful event, a gestalt.

The Helmholtz principle, conversely to classical statistical methods, does not require a precise modelization of the observed phenomenon. In fact it coarsely models a statistical background that represents the absence of significant events. 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 signals.

3. Audiovisual Gestalts

As underlined at the end of previous section, the audiovisual configuration that we want to detect has to be defined such that it depicts a perceptually meaningful structure. The starting observation here is that visual signals are mainly made of moving regions surrounded by contours with high geometrical content. An image sequence can thus be decomposed into 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 through time, we define video structures that obey Gestalt principles. In particular, sets of individual pixels are grouped together according to proximity, similarity and common motion, three of the basic Gestalt laws (see Sec. 1).

The video representation algorithm was developed in [5], and it was adopted for the analysis of multi-modal signals in [13]. It is briefly introduced in the next section, while in Sec. 3.2 the audio representation method is described and in Sec. 3.3 meaningful audiovisual events are defined.

3.1. Video Representation

The image sequence is decomposed into a set of video atoms which represent salient geometric video components and track their temporal transformations. The use of geometric video decomposition has two main advantages. When considering time-evolving image structures in fact, we use dynamic features with a true geometrical meaning. Moreover, sparse 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 video sequences, which have a very high dimensionality.

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. Assuming that this high-pass image $I(\vec{x})$ can be approximated with a linear combination of functions $G_\gamma(\vec{x})$ (called *atoms*) retrieved from a redundant dictionary $\mathcal{D}_\mathcal{V}$ of 2-D atoms, we can write:

$$I(\vec{x}) \approx \sum_{\gamma_j \in \Gamma} c_{\gamma_j} G_{\gamma_j}(\vec{x}) \,, \tag{1}$$

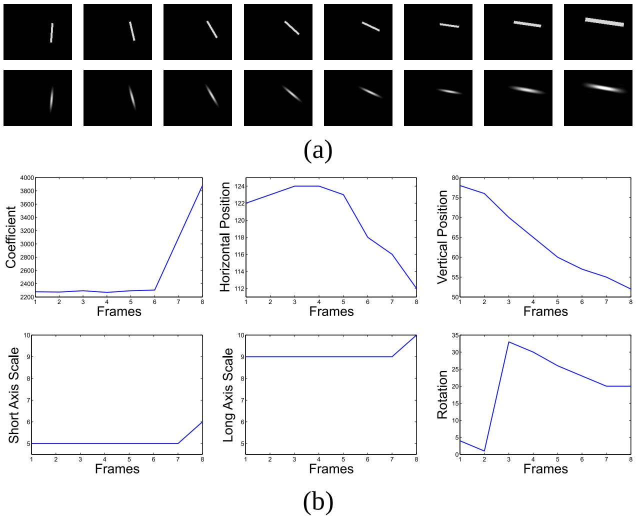
where j is the summation index, c_γ corresponds to the coefficient for every atom G_γ and Γ is the subset of selected atom indexes from dictionary $\mathcal{D}_\mathcal{V}$. The codebook $\mathcal{D}_\mathcal{V}$ is built by applying a set of geometric transformations to a mother function $G(\vec{x})$, in order to generate an overcomplete set of primitives spanning the input image space. The considered transformations are anisotropic scaling s_1 and s_2, translations t_1 and t_2 and rotation θ. The generating function G should represent well edges and thus, it should behave like a smooth scaling function in one direction and should approximate the edge along the orthogonal one. We use an edge-detector atom that is a Gaussian along one axis and the first derivative of a Gaussian along the perpendicular one. The decomposition of $I(\vec{x})$ on an overcomplete dictionary is not unique. Because of its simplicity, in this paper we use Matching Pursuit (MP) [12], an iterative greedy algorithm that selects the element of the dictionary that best matches the signal at each iteration.

We consider an approach where 2-D spatial primitives G_γ obtained in the expansion of a reference frame of the form of Eq. 1 are tracked from frame to frame. 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

$$I_{t+1}(\vec{x}) = \sum_{\gamma_j \in \Gamma} F_t^{\gamma_j} \cdot (c^t_{\gamma_j} G^t_{\gamma_j}(\vec{x})) \,, \tag{2}$$

where F_t represents the set of transformations F_t^γ of all atoms that approximate each frame. A MP-like approach similar to that used for the first frame is applied to retrieve the new set of $G_\gamma^{t+1}(\vec{x})$ (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 [5], to which the interested reader is referred.

A cartoon example of the used approach can be seen in Fig. 1, where the approximation of a simple synthetic object by means of a single video atom is performed. Fig. 1(a) shows the original sequence (top row) and its approximation composed of a single geometric term (bottom row). Fig. 1(b) depicts the parametric representation of the sequence: we find the temporal evolution of the coefficient c_γ and of the position, scale and orientation parameters. The MP video representation provides a parametrization of the signal which concisely represents the image geometric structures *and* their temporal evolution.



(a)

(b)

Figure 1. (a) [Top row] Original synthetic sequence made by a moving line. [Bottom row] Approximation using one video atom. (b) Parameter evolution of the atom. From left to right and from up down: coefficient c_γ, horizontal position t_1, vertical position t_2, short axis scale s_1, long axis scale s_2, rotation θ.

3.2. Audio Representation

In this work, we want to detect synchronous audio-video events. In this context an interesting audio event is the presence of a sound. Thus, we need an audio feature which simply allows to assess the presence or not of an acoustic event. We consider here an estimate of audio energy contained per frame. To compute such an estimate, we exploit again the properties of signal representations over redundant dictionaries using MP [12]. 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 MP over a redundant dictionary $\mathcal{D}_\mathcal{A}$ of unit norm atoms. The codebook $\mathcal{D}_\mathcal{A}$ is generated by scaling, translating in time and modulating in frequency a generating function $g(t) \in L^2(\mathbb{R})$. We use here a dictionary of Gabor atoms, *i.e.* the function $g(t)$ is a normalized Gaussian window, which has been chosen for its optimal time-frequency localization [12].

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

$$a(t) \approx \sum_{\omega_i \in \Omega} c_{\omega_i} g_{\omega_i}(t), \tag{3}$$

where c_{ω_i} are the coefficients and Ω 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 easily derived from its MP decomposition [12]. 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 [13]. An example of one function $f_a(t)$ is shown in Fig. 2(b).

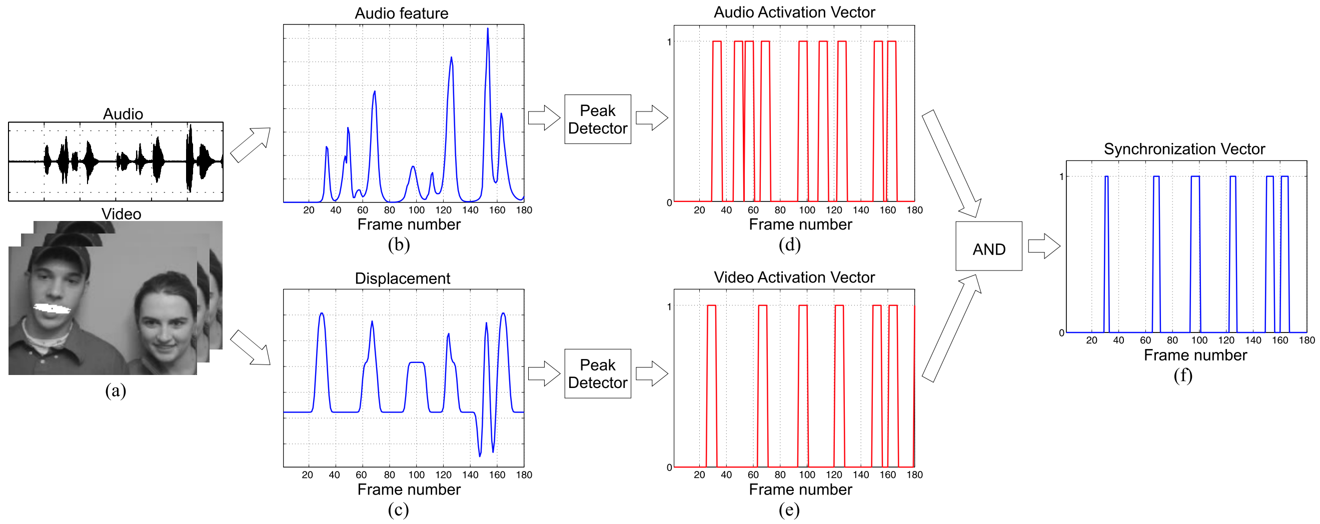


Figure 2. Scheme of the proposed audiovisual fusion criterion. Starting from the audiovisual sequence (a), we compute the audio feature $f_a(t)$ (b), and the displacement feature for a video atom representing the speaker's mouth (c). The two features exhibit a remarkable synchrony. From these signals 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 created computing the logical AND between the audio-video activation vectors (f).

3.3. Meaningful Audiovisual Events

The audio feature $f_a(t)$ basically estimates the average energy present in the audio signal $a(t)$. The output of the MP video algorithm, instead, is a set of atom parameters describing the temporal evolution of the video features. From the positions, we can compute the displacement of each video atom and thus estimate the movement of important visual structures. For each video atom we compute the absolute value of the displacement as $d = \sqrt{t_1^2 + t_2^2}$, with t_1 and t_2 respectively horizontal and vertical positions of the atom. In order to be more easily compared to the audio feature and to filter out small spurious movements, we convolve the video feature d with a Gaussian kernel, obtaining a smooth function like the one depicted in Fig. 2(c).

We have now one audio feature and N video features describing the movement of relevant visual features, where N is the number of atoms used to represent the video. Each of these variables has the same number of samples T, since we downsample $f_t(a)$ that has a higher temporal resolution. Peaks in these signals suggest the presence of an event. In the video case, it can be the movement with respect to a certain equilibrium position (*e.g.* lips opening and closing). For the audio, a peak indicates the presence of a sound. The temporal proximity of such audio and video peaks suggests the presence of a gestalt reflecting two expressions of the same 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 which models delays and uncertainty. 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 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, denoted as *synchronization vectors* $s_i(t)$. The vectors s_i equal 1 at time instants at which both audio and video atoms are active and 0 otherwise. Thus, the number of 1 in the vector indicates the degree of synchronization between the audio-video pair. Fig. 2 summarizes the construction of one synchronization vector $s_i(t)$.

4. Detection of Audiovisual Events

Once synchronization vectors are available, we need a method to select those vectors (and thus those audiovisual structures) associated to *meaningful* audio-video pairs. We want to do that in an automatic way, tuning as less parameters as possible. For each video atom we have one synchronization vector $s_i(t)$. Suppose that we observe a synchronization vector of length n (*i.e.* that is built over an observation 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 answer these questions using the Helmholtz principle.

We first have to define the background *a contrario* model which corresponds to the absence of correlated audiovisual events. In this case the observations $s_i(t)$ are considered as independently, identically distributed random variables. Since the general form of their distribution is unknown (anyway, it is not reasonable to assume that a single distribution could account for all the sequences), the empirical distribution is considered [3]. Integrating this distribution yields the function $P_s(X)$, where X is a random variable distributed according to the empirical distribution of the observed values $s_i(t)$ (with $i = 1, \dots, N$).

Let A be a video atom with corresponding synchronization vector s_{A} of length n, and let k be the number of points at which s_{A} assumes value 1. Let us define the event $E =$ *"At least k points of a synchronization vector s_{A} of size n keep a value equal to 1"*. Thus, according to the background model, the probability of the event E, $P(E)$,

$$P(E) = \mathcal{B}(k, n, P_s(s_{\mathsf{A}} = 1)) \,, \tag{4}$$

where $P_s(s_{\mathsf{A}} = 1)$ is directly deduced from $P_s(X)$ and $\mathcal{B}(k, n, p)$ is the tail of a binomial distribution:

$$\mathcal{B}(k, n, p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \,. \tag{5}$$

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

Definition 1 *For a given atom A with corresponding synchronization vector s_{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:*

$$NFA(\mathsf{A}) = N \cdot \mathcal{B}(k, n, P(s_{\mathsf{A}} = 1)) \,, \tag{6}$$

with N number of tests. In this context N is the number of video atoms used for the decomposition of the sequence.

An atom A is said to be ε-meaningful if $NFA(\mathsf{A}) \le \varepsilon$.

It is easy to demonstrate that the expected number of ε-meaningful video atoms in a sequence, according to the *a contrario* model, is less then ε and that the number k of matching points required for a vector to be significative depends on the logarithm of ε and N [3]. This means that the detection results are robust to variations of those values.

The value of ε controls the number of false detections. Setting ε equal to 1, as in [2], means that the expected number of false detections 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 MP video algorithm exploits the correlation between neighboring atoms [5, 13]. Because of that, some video atoms exhibit NFA smaller then $\varepsilon = 1$, even without being correlated with the audio. One solution is that of considering a smaller value of ε, as it is done in [3] where $\varepsilon = 1/10$.

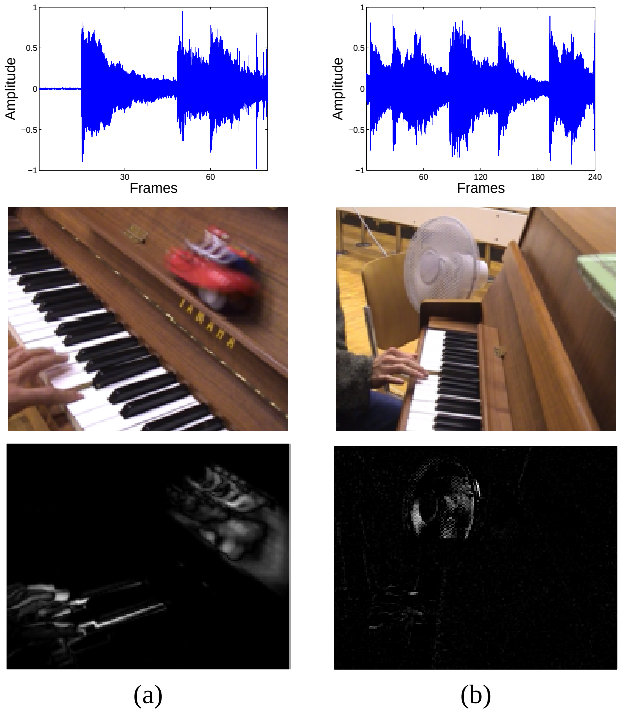


Figure 3. Test sequences Piano 1 (a) and Piano 2 (b). [Top] Audio tracks, [Middle] sample frames, [Bottom] corresponding dynamic pixels: gray-levels represent the absolute value of the difference between the luminance components of two successive frames.

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 ε (smaller than 1), keeping the solution which is more localized in space. By doing that, we basically do not fix a detection threshold. Instead, we browse a set of solutions and we chose the most suitable one. In practice, we consider a set of thresholds ε_i uniformly spaced in a logarithmic scale between ε_{MIN} and 1. For each value ε_i, we obtain a set of video atoms G_i for which $NFA(\mathsf{A}) \le \varepsilon_i$, with $\mathsf{A} \in G_i$. For each group G_i, the variances along the horizontal (var$_x$) and vertical positions (var$_y$) are computed and the maximum value $V_{G_i} = \max\{\text{var}_x(G_i), \text{var}_y(G_i)\}$ is kept. Our solution G^* is the set of atoms which exhibits the smallest variance V_{G^*}.

5. Experiments

We show here how the proposed framework is used to locate the source of an audio signal in real video sequences. The first test involves two clips, denoted as Piano 1 and Piano 2. They both show a hand playing piano while some distracting visual and acoustic noise is present. Sample raw frames of the sequences are shown in Fig. 3. In Piano 1 a toy car is passing through the scene, while in Piano 2 a ventilator is on and it is moving from left to right. These examples have been chosen to demonstrate the robustness of the proposed algorithm to audio distractors, thanks to the denoising properties of the audio MP decomposition, and to video distractors both of constant velocity (Piano 1) and oscillating (Piano 2). The clips were recorded at 25 frames/sec (fps) at a resolution of 144×180 pixels and only the luminance components were considered. The soundtrack was collected at 44 kHz and sub-sampled to 8 kHz.

Image sequences are represented with 40 video atoms, while the audio track is decomposed using 1000 Gabor atoms. 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 set of meaningful atoms G^* is selected using $\varepsilon_{MIN} = 10^{-5}$ and the thresholds $\varepsilon_i = \{10^{-5}, 10^{-4.5}, 10^{-4}, \ldots, 1\}$. 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.

In order to take into account the dynamics of the scene, a sliding observation window over which the synchronization vectors are computed has to be used. A window of 60 frames length is used to detect the video atoms that are more correlated with the audio following the procedure described in Sec. 3 and Sec. 4. The observation window is then 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, the choice of none of the parameters results to be critical.

Fig. 4 shows resulting sample frames of the algorithm run on the sequence Piano 1. In white we highlight the footprints of the video atoms which are found to be more correlated with the soundtrack. The player's fingers are detected as sound sources. The moving toy car introduces a considerable distracting motion (see Fig. 3 (a)) and a non-negligible acoustic noise. However, it is filtered out by the cross-modal localization algorithm. Fig. 5 shows the same type of results for clip Piano 2. It is interesting to remark that in this case the visual distractor (the ventilator) does not have a constant velocity as in the previous case, but it is oscillating in the background. This results in peaks in the video activation vectors representing the ventilator's edges. However, these oscillating structures are not detected as correlated with the audio, since they are not synchronous with the audio activation peaks.

A second set of experiments has been carried out on four sequences taken from the CUAVE database [15], in order to test the proposed algorithm in a multi-modal speaker localization task and to compare its performances to those of existing methods. 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 soundtrack was collected at 44 kHz and sub-sampled to 8 kHz. The setting of the experiments is the same described above and all the parameters keep the same values. The test clips are referred to with the names they have on the CUAVE dataset, *i.e.*g19, g20, g21, g22. The sequences involve two persons arranged as in Fig. 6 taking turn in reading series of digits. Fig. 6 shows the results for sequence g22. In the first sample frame the left person is speaking, while in the second the right one is. The sequence is non-trivial, since the left person mouths the digits which are being uttered by the right speaker. The algorithm is able to correctly localize the mouth and the chin of the current speaker. It is interesting to remark how video atoms correlated with the sound shift from one speaker's mouth to the other, handling the dynamics of the scene.

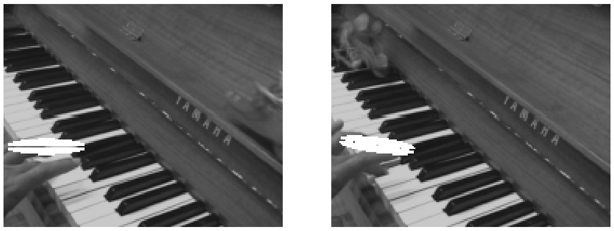


Figure 4. Results of the proposed algorithm run on the clip Piano 1. The most correlated atoms, highlighted in white, represent the player's fingers. The moving toy car is not detected.

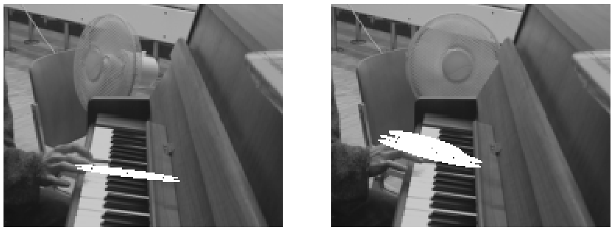


Figure 5. Results for the sequence Piano 2. The correlated atoms, highlighted in white, are on the player's fingers and the piano keys. The oscillating ventilator is not detected.

In order to quantify the accuracy of the proposed algorithm, we have manually labelled the center of the speaker's mouth in the test sequences. The active speaker's mouth is considered to be correctly detected if the position of the most correlated video atom falls within a circle of diameter D centered in the labelled mouth center. If more than one atom is chosen, an atoms' centroid is estimated whose position on the image plane is given by the average of the single atoms coordinates. Since correlated atoms are detected every 20 frames, mouth labels are placed with this same frequency throughout each sequence, and performances are

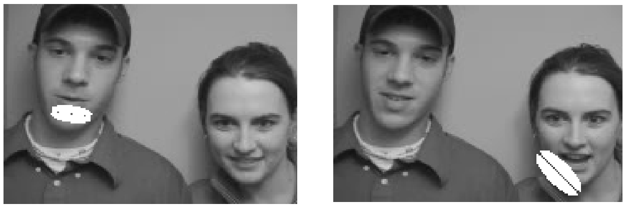


Figure 6. Results for sequence g22: in the first sample frame the left person is speaking, while in the second the right one is. The most correlated 3-D atoms are highlighted in white. The mouth and the chin of the correct speaker are detected.

thus evaluated at test points distant 20 samples one from the other. The value of the diameter D is set to 50 pixels. This value has been chosen so that we can compare the results with those presented in [14] and [13].

Nock and colleagues [14] propose a method to detect the mouth of the speaker founding the image zone over which the mutual information between audio and video features is maximized. As in our algorithm, in [14] mutual information values are estimated using a sliding time window of 60 frames that is shifted in time with steps of 30 frames. The goodness of the detection is assessed using the criterion that we use here, with the only difference that in [14] the speaker's mouth is considered to be correctly located if it is placed within a *square* of 200 × 200 pixels centered on the manually labelled mouth center. 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. However, we must note that the test clips used in [14] could not exactly coincide with those used in this paper, since the original sequences have been cropped in both cases. In contrast, the results presented in [13] are obtained using exactly the same test sequences. The algorithm in [13] served as inspiration for this work and thus they have many common points. The main difference is that in [13] the video structures that are detected are simply those exhibiting the highest degree of synchrony with the audio. Here in contrast, the audiovisual fusion problem is defined as a gestalt detection problem, which allows to automatically set a saliency detection threshold.

Table 1 summarizes the results obtained for the three methods in term of percentage of test points at which the speaker's mouth is correctly detected. Note that there could be no perfect coincidence between the test sequences used in [14] and those used in [13] and here, thus the results for Nock's algorithm should be considered only as indicative. As already shown in [13], the algorithm by Monaci *et al.* in general improves the results obtained by Nock and colleagues. The proposed method obtains detection performances similar to those of Monaci's algorithm, slightly improving previous results for sequence g21. We want to underline again that in contrast to previous methods, we do not simply seek for the video region that maximizes the correlation with the audio, but more generally we look for image zones whose synchrony with the audio are above a saliency threshold. This threshold does not require to be tuned, since a set of meaningful thresholds is fixed in advance and the one giving the most suitable solution is adopted.

Clip	Nock[14]*	Monaci[13]	**Proposed**
g19	41	87	**87**
g20	93	93	**93**
g21	79	78	**81**
g22	79	87	**87**

Table 1. Results expressed in percentage of correct detections. *These values should be considered only indicative (see text).

The audio-video gestalts that are detected have a high semantic meaning. This allows to extract and manipulate these structures in a simple and intuitive way. For example, it is possible to reconstruct the scene using only those video atoms that are consistent with the audio track by simply encoding the video sequence with 3-D atoms that are close to the detected sound source. Fig. 7 shows sample raw frames of clip g20 and their reconstruction obtained by summing to the low-pass images those video atoms that are closer than $R = 80$ pixels to the estimated sound source. The reconstructed images can be seen as *audiovisual key frames* that focus on the sound source at a given time instant. Moreover, in a compression application scenario, a sequence can be selectively encoded using only video atoms associated with the soundtrack, saving bits for the coding while keeping the salient information about the scene.

6. 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 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 and effective audiovisual fusion criteria and demonstrates that temporal proximity between audiovisual events is a key ingredient for cross-modal integration of information. The proposed method exhibits robustness to significant audio-video distractors. In addition, the considered audiovisual structures have a high semantic role and can be easily extracted and manipulated.
- *The algorithm naturally deals with dynamic scenes.*

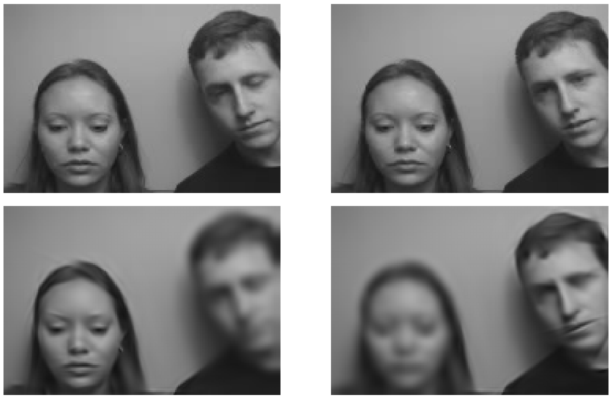


Figure 7. Sample raw frames of clip g20 [Top] and reconstruction using only video atoms close to the estimated sound source [Bottom]. On the first sample the left person is speaking while on the second one the right person is speaking.

- *There is no parameter to tune*. All parameters are fixed and from informal tests the algorithm performances turn out to be 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 40 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 (*e.g*. size and orientation of video structures) that can be exploited at successive processing stages.

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 [9].

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 the proposed algorithm to other types of multi-modal 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 and to Frédéric Cao for reading this paper and introducing us to Gestalt theory. This work is supported by the Swiss National Science Foundation through the IM.2 NCCR.

References

[1] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. on PAMI*, 25(7):828–836, 2003.

[2] F. Cao. Application of the Gestalt principles to the detection of good continuations and corners in image level lines. *Computing and Visualization in Science*, 7:3–13, 2004.

[3] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.

[4] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Trans. on PAMI*, 25(4):508–513, 2003.

[5] O. Divorra Escoda. *Toward Sparse and Geometry Adapted Video Approximations*. PhD thesis, EPFL, June 2005. [Online] Available: http://lts2www.epfl.ch/.

[6] J. Driver. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, 381:66–68, 1996.

[7] J. W. Fisher III and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, 6(3):406–413, 2004.

[8] J. Hershey and J. Movellan. Audio-vision: Using audiovisual synchrony to locate sounds. In *Proc. of NIPS*, 1999.

[9] P. Jost, P. Vandergheynst, and F. P. Tree-based pursuit: Algorithm and properties. *IEEE Trans. on Signal Processing*, in press, 2006. [Online] Available: http://lts2www.epfl.ch/.

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

[11] E. Kidron, Y. Schechner, and M. Elad. Pixels that sound. In *Proc. of CVPR*, pages 88–95, 2005.

[12] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12):3397–3415, 1993.

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

[14] H. J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: an empirical study. In *Proc. of the 10^{th} ACM Int. Conf. on Multimedia*, 2002.

[15] 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*, (11):1189–1201, 2002.

[16] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004.

[17] M. Slaney and M. Covell. FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Proc. of NIPS*, 2000.

[18] P. Smaragdis and M. Casey. Audio/visual independent components. In *Proc. of ICA*, pages 709–714, 2003.

[19] 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*, 158:252–258, 2004.