Audio-driven Nonlinear Video Diffusion

Anna Llagostera Casanovas* and Pierre Vandergheynst
Signal Processing Laboratory (LTS2)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Station 11, 1015 Lausanne, Switzerland
e-mail: anna.llagostera@eecs.qmul.ac.uk,pierre.vandergheynst@epfl.ch
phone: +41 21 693 26 01, fax: +41 21 693 76 00

***Abstract*—In this paper we present a novel nonlinear video diffusion approach based on the fusion of information in audio and video channels. Both modalities are efficiently combined into a diffusion coefficient that integrates the basic assumption in this domain, i.e. related events in audio and video channels occur approximately at the same time. The proposed diffusion coefficient depends thus on an estimate of the *synchrony* between sounds and video motion. As a result, information in video parts whose motion is not coherent with the soundtrack is reduced and the sound sources are automatically highlighted. Several tests on challenging real-world sequences presenting important auditive and/or visual distractors demonstrate that our approach is able to prevail regions which are related to the soundtrack. In addition, we propose an application to the extraction of audio-related video regions by unsupervised segmentation in order to illustrate the capabilities of our method. To the best of our knowledge, this is the first nonlinear video diffusion approach which integrates information from the audio modality.**

***Index Terms*—Audio-visual processing, linear/nonlinear diffusion, graph cut segmentation**

I. Introduction

The perception that we have about the world is influenced by elements of diverse nature. Indeed humans tend to integrate information coming from different sensory modalities to better understand their environment. In the audio-visual domain for example, the listener can exploit the correspondence between speaker lips movements and the produced sounds to better understand speech, especially in adverse environments [1–3]. The speech recognition task is thus facilitated by the integration of acoustic and visual stimuli. Following this observation, scientists have been trying to combine different research domains. Nowadays it is possible to use the video information to improve results in the audio domain for applications such as speech recognition [4, 5], speech enhancement [6, 7] and sound source separation [8–10]. Other methods try to assess coherence between both modalities to track or locate sound sources in the video signal [11–17]. Some approaches go one step beyond and try to separate the scene into audio-visual structures, each of them composed by a visual part and the associated soundtrack [18–20]. All these applications can then be used for automatic management of videoconferences, automatic speaker recognition [21], indexing and segmentation of multimedia data [22], and robotics [23].

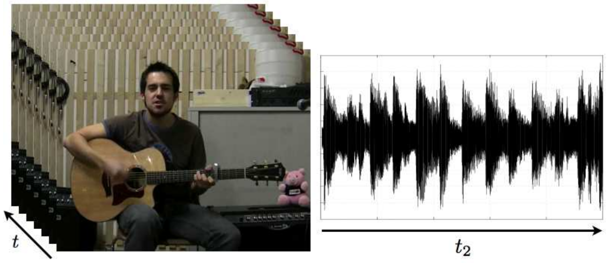


Fig. 1. Example of a 3D video signal [left] and the corresponding 1D audio signal [right]. The temporal axis of each modality has a different resolution.

Fig. 1 depicts the typical baseline in audio-visual analysis. We have a three-dimensional video signal recorded with *one video-camera* and the corresponding one-dimensional audio signal captured by *one microphone*. Notice that here we consider the simplest audio-visual configuration, which does not include microphone arrays. As shown in Fig. 1, audio and video signals share a temporal axis, but the resolution of this axis is different. Typically, we have much more audio samples than video frames since the sampling rate of the audio signal is much higher. Then the challenge lies in efficiently combining the information in both channels.

Many approaches in this domain first define features for each modality such as the energy [17, 24, 25] or Mel-Frequency Cepstral Coefficients (MFCC) [12, 13, 26, 27] for the audio, and pixel intensities [15, 24, 28] or temporal variations [13, 17, 27] for the video. Then, they use these representations in a fusion step, whose objective is to assess the synchrony between both modalities using canonical correlation analysis [12, 17, 20] or through the estimation of the joint densities of audio and video features [13, 14, 24, 26, 28]. Most of those methods are based on pixel behavior, which makes them vulnerable to visual noise. Furthermore, they do not ensure video spatial coherence. Other methods propose to decompose each modality [19] or both modalities at the same time [29] over redundant dictionaries of signals. That makes the fusion step more intuitive since they deal with audio and video *structures* and the signals can be expressed as a sum of a small number of functions. As a result, the computational cost for the audio-visual fusion step is much smaller than in pixel-based methods [13, 24, 26, 28]. However, the decomposition

Web page: http://lts2www.epfl.ch
This work was partly funded by the Swiss NFS through grant number 200021-117884 and by the EU Framework 7 FET-Open project FP7-ICT-225913-SMALL: Sparse Models, Algorithms and Learning for Large-Scale data.

*Anna Llagostera Casanovas is now with the Vision group at Queen Mary University of London, London, UK.

of audio and video signals into such meaningful structures is time consuming.

As discussed before, approaches in audio-visual analysis try to assess the *synchrony* between audio and video channels in order to extract information about the observed scene. Thus, in most applications only the video parts that are related to the soundtrack are used. For example, speech recognition only needs the region around the mouth, and approaches in sound source localization search for regions moving coherently with the sounds. The remaining video information (i.e. background and video structures not related to the soundtrack) is superfluous and not necessary for those audio-visual applications. However, identifying a mouth or discriminating audio-related motion from distracting motion involves a significant amount of computational cost.

The aim of this work is to simplify audio-visual sequences by eliminating most of this non-relevant video information through a cheap and fast procedure. After Perona and Malik's preliminary work in [30], nonlinear diffusion (also called anisotropic diffusion) has been proven to be a useful tool for the *selective* removal of information in a given signal. This technique has been successfully applied to image denoising, restoration and edge detection [31–34]. Furthermore, the flexibility in the design of the diffusion coefficient (which controls the intensity of the diffusion at each point of the signal) makes it applicable to a great variety of problems.

The main contribution of our approach is the definition of an audio-visual diffusion coefficient, which integrates the basic assumption in this domain and represents a natural way to combine audio and video modalities. The proposed diffusion coefficient is a function of the *synchrony* between audio energy and video motion at each point of the video domain. As a result, our diffusion procedure removes information in parts of the video signal whose motion is not coherent with a synchronously recorded audio track, while preserving *regions* that are useful for audio-visual applications.

In summary, the main strengths of our approach are:

1) Our method can handle all kind of audio-visual sources since it is based on a general assumption, i.e. synchrony between related events in audio and video channels.
2) The 3D characteristic of the diffusion process implicitly brings spatio-temporal coherence to our approach, by prevailing regions instead of pixels.
3) The proposed method can deal with multiple audio-visual sources because video structures in different locations are treated independently, i.e. we do not need to chose a region to preserve over the rest.

To the best of our knowledge, this is the first nonlinear video diffusion approach which integrates information from the audio modality.

The paper is structured as follows. Sec. II presents the proposed model for audio-based nonlinear video diffusion. Sec. III explains the discrete implementation of our method and presents a stopping criterion for the diffusion process. In Sec. IV we detail the audio and video features used in our approach. Sec. V introduces a quantitative measure of our method's efficiency. In Sec. VI we show the results when analyzing challenging audio-visual sequences. Sec. VII proposes a simple application of our method to the unsupervised extraction of audio-related video regions. Finally, in Sec. VIII achievements and future research directions are discussed. Partial results were presented in [35].

II. Audio-based Video Diffusion

Our diffusion model is inspired by the variant of the classic Perona-Malik model [30] that Catté et al. proposed in [32]. This nonlinear diffusion approach based on partial differential equations (PDEs) has been demonstrated to provide good results in the previously mentioned applications. Sec. II-A recalls the main principles of PDE-based diffusion and Sec. II-B describes the proposed audio-visual diffusion coefficient.

A. PDE-based Diffusion

Let us consider a 3D video domain Ω with boundary $\Gamma := \partial\Omega$ and let a video signal v be represented by a mapping $f \in \mathrm{L}^\infty(\Omega)$. Then, a general continuous model for anisotropic diffusion filters is represented by the following boundary value problem:

$$\partial_\tau v = \operatorname{div}(D\nabla v) \quad \text{on} \quad \Omega \times (0, \infty)\,, \tag{1}$$
$$v(\mathbf{x}, 0) = f(\mathbf{x}) \quad \text{on} \quad \Omega\,, \tag{2}$$
$$\langle D\nabla v, \mathbf{n}\rangle = 0 \quad \text{on} \quad \Gamma \times (0, \infty)\,. \tag{3}$$

Here D is a positive definite *diffusion coefficient*, τ refers to the diffusion time, $\mathbf{n}$ denotes the outer normal, $\mathbf{x} = (x, y, t)$ are the 3D video coordinates, $\langle ., . \rangle$ is the Euclidean scalar product on $\mathbb{R}^3$, $\operatorname{div}(\cdot)$ and ∇ denote, respectively, the divergence and the gradient operators with respect to the space variables. Notice that τ is used for the diffusion time and t for the temporal axis of the video signal. This notation will be kept throughout the paper.

The diffusion equation in (1) belongs to a general class of equations satisfying the *maximum principle*. The principle states that all the maxima of a solution of Eq. (1) for diffusion times $\tau \in [\tau_0, \tau_1]$ are to be found on the boundary Γ or at $\tau = \tau_0$ provided that the diffusion coefficient D is positive. Since our boundary problem is also composed of Eq. (3), the diffusion is 0 across the boundary Γ and the maxima can only belong to the original video (initial condition at $\tau = \tau_0$). A proof of the maximum principle can be found in [36]. In practice, this is a very important property since the principle prevents the creation of new local extrema when applying the diffusion process to any function v.

For a deeper understanding of PDE-based diffusion, please refer to the works in [31, 33].

B. Audio-Visual Diffusion Coefficient

We propose the following diffusion coefficient D:

$$D(\mathbf{x}, \tau) = g(|s_\sigma(\mathbf{x}, \tau)|^2)\,, \tag{4}$$

where $g(\cdot)$ is a function that determines the intensity of the diffusion process at each point of the video volume and s_σ

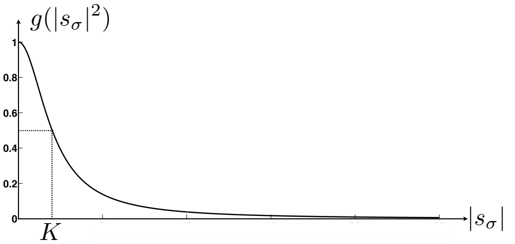


Fig. 2. Shape of the function $g(\cdot)$ in Eq. (6), which determines the value of the diffusion coefficient according to the audio-visual synchrony s_σ.

is a regularized measure of the synchrony between events in audio and video channels, which is defined as

$$s_\sigma(\mathbf{x}, \tau) = (a(\mathbf{x})\partial_t v(\mathbf{x}, \tau)) * G_\sigma(\mathbf{x}) \,. \qquad (5)$$

In this expression, G_σ is a 3D Gaussian of variance σ^2, $\partial_t v$ is the temporal derivative of the video signal, and $a(x, y, t) = a(t)\ \forall x, y$ represents the energy of the audio channel at time t (notice that the audio feature does not depend on the spatial coordinates x and y). Thus, the *audio-video synchrony* s_σ evaluates the coherence between both channels by combining audio energy and video motion at each point $\mathbf{x}$ of the video volume. According to Eq. (5), $|s_\sigma|$ is high when an important acoustic event matches a relevant pixel motion while its value is close to zero in the rest.

The convolution with a Gaussian G_σ in Eq. (5) makes our audio-visual synchrony measure s_σ much more robust to visual and acoustic noise and ensures spatio-temporal coherence to our method. Furthermore, this procedure was used by Catté et al. in [32] in order to regularize the nonlinear diffusion problem presented by Perona and Malik in [30], whose formulation is similar to ours. In all experiments the regularization parameter is fixed to $\sigma = 1$. This value has been shown in [34] to be sufficient for a large interval of noise variances when the noise in neighboring pixels is uncorrelated and the grid size is one.

Let us now discuss the shape of the function $g(\cdot)$ in Eq. (4). As explained before, we want a linear diffusion process to take place in spatio-temporal regions with low audio-visual synchrony. In addition, the diffusion coefficient D should be close to 0 in points with high $|s_\sigma|$ in order to stop there the diffusion. Thus, $g(\cdot)$ should be a non-negative monotonically decreasing function with $g(0) = 1$, since the diffusion coefficient D has to be positive. An appropriate shape for $g(\cdot)$ can then be the function proposed by Perona and Malik in [30] (see Fig. 2):

$$g(|s_\sigma|^2) = \frac{1}{1 + \frac{|s_\sigma|^2}{K^2}} \,. \qquad (6)$$

The value of the constant K acts as a threshold: points where $|s_\sigma| < K$ are strongly affected by linear diffusion (Gaussian blurring) while those points where $|s_\sigma| > K$ are least diffused. Appropriate values for this parameter are discussed in the experiments section.

We can now analyze qualitatively the behavior of the proposed audio-visual diffusion process given the diffusion coefficient defined in Eq. (4). First of all, the diffusion coefficient is maximal and constant to $D(\mathbf{x}, \tau) = 1$ in video *regions* where $s_\sigma = 0$, that is:

1) Static video regions (video inactivity).
2) Silent time slots (audio inactivity).
3) Situations where the visual motion is not synchronous with the appearance of sounds (audio-video incoherence).

Inside these regions, the diffusion coefficient is constant to 1, the diffusion equation in (1) becomes the heat equation ($\partial_\tau v = \Delta v$) and the region is homogeneously diffused. Out of those regions, the diffusion coefficient D becomes smaller and the diffusion process is stopped. In fact, the larger is $|s_\sigma|$ the lower is the level of diffusion that a pixel experiences. In addition, the nature of linear 3D diffusion together with the regularization with a Gaussian G_σ in Eq. (5) implicitly bring spatial coherence to our approach by prevailing structures over pixels. Notice that the diffusion coefficient $D \approx 1$ in a pixel that is surrounded by pixels with low audio-visual synchrony $|s_\sigma|$, independently of the synchrony of the pixel itself. Thus, only spatio-temporal *regions* whose movement is coherent with the soundtrack are preserved.

As a summary, we are performing a *nonlinear* diffusion over a 3D volume (the video signal) which is controlled by a diffusion coefficient D that depends on the coherence of audio and video signals. The proposed diffusion process leads to the blurring of the visual structures that are not relevant for audio-visual analysis while keeping a good resolution in the rest. Thus, in the resulting video signal the possible sound sources in the scene are automatically highlighted. Some examples of this behavior can be seen in Fig. 3 (b), where the hand that is playing the piano (audio-visual source) is better preserved than the other elements in the scene (e.g. piano brand in the top sequence).

Other strategies could be considered in the definition of an audio-visual diffusion process. For example we could use other approaches such as [14, 17, 18, 29] to estimate in a first stage the location of the sound sources in the image. Then, it would be possible to define a diffusion coefficient which does not change through time and simply blur video regions that are far from the estimated position of the source. An approach in this direction can be found in [38], were the authors use audio-video analysis to encode regions close to the source's location with more quality than other regions in the video. However, our purpose is to preserve *only* the video structures whose motion is related to the soundtrack. Approaches in audio-visual speech recognition require the lips' shape and movements (maybe also the mouth region) but not the speaker's eyes for example. After localizing the sound source with other methods we would still need to define the region to preserve, and by using a proximity criterion we would obtain a high-resolution region with a circular shape and an arbitrary radius. An example of the resulting signal after such a "Localize & Diffuse" approach can be observed in Fig. 3 (c). In this case, using a fixed value for the radius that defines the region to preserve might not be appropriate, since it could lead to the removal of important information in some cases and to the preservation of irrelevant details in other sequences. In contrast, by using our method only edges whose motion is coherent with the soundtrack are preserved,

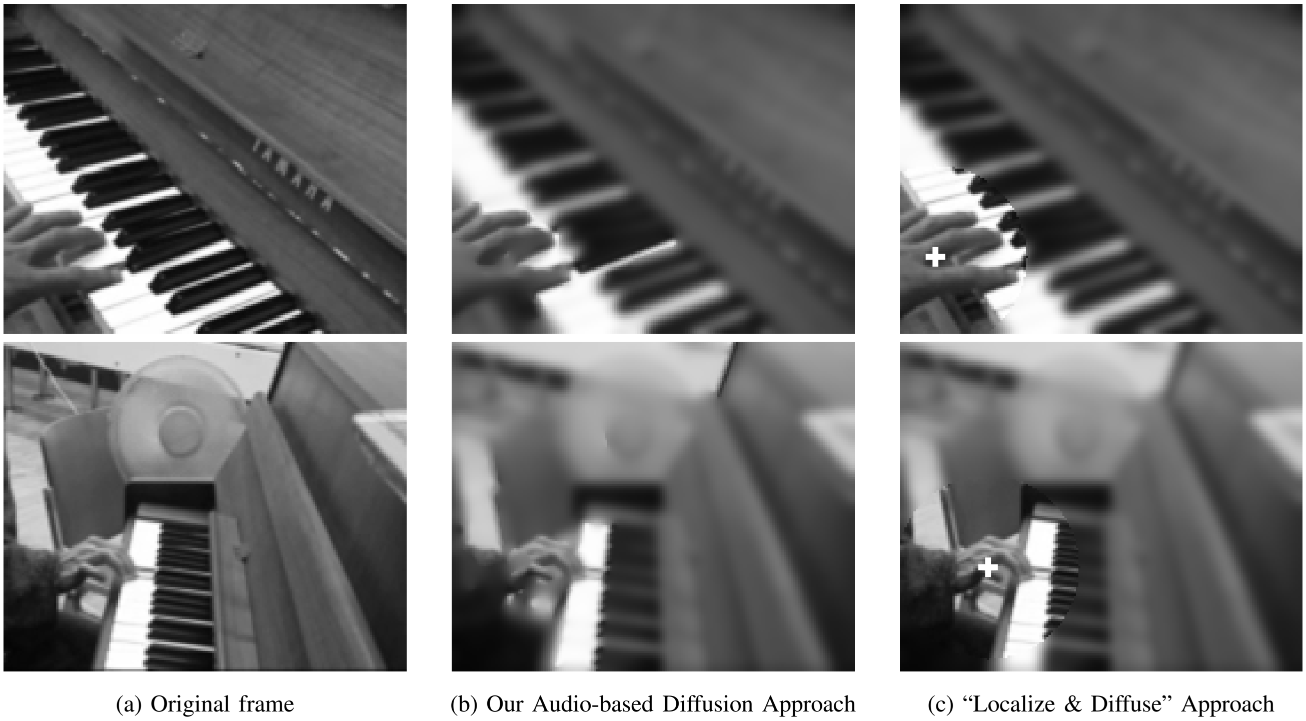

(a) Original frame (b) Our Audio-based Diffusion Approach (c) “Localize & Diffuse” Approach

Fig. 3. Results when applying the proposed audio-based video diffusion approach (b) and a method based on a “Localize & Diffuse” strategy (c), which first localizes the audio-visual source and then applies a Gaussian blurring to the pixels further than 40 pixels from the source position. White crosses in (c) represent the source position (hand), which is manually fixed for this visualization. Sequences belong to the audio-visual source localization method in [37].

and the size and shape of the source(s) does not need to be specified in advance.

III. Discretization and Stopping Criterion

A. Discretization

We have previously presented the continuous model for the audio-based nonlinear video diffusion. The discretization of the proposed approach by means of finite differences can be found in [35].

Let $v^n_{i,j,k}$ be the value of v at location $(i\Delta x, j\Delta y, k\Delta t)$ and diffusion time $n\Delta\tau$. Here Δx, Δy and Δt are the grid spacing used in the discretization of the video dimensions, while $\Delta\tau$ is the grid spacing used for the diffusion time discretization. In our case, the pixel size is chosen as the unit of reference in all spatio-temporal dimensions: $\Delta x = \Delta y = \Delta t = 1$. The discretization scheme in [35] satisfies the maximum and minimum principle (whose importance was discussed in Sec. II-A) for a choice of $\Delta\tau \in [0, 1/6]$. Thus, if we define the maximum and the minimum of the neighbors of $v_{i,j,k}$ at iteration n as $v_M = \max\{(v, v_l)^n_{i,j,k}\}$ and $v_m = \min\{(v, v_l)^n_{i,j,k}\}$, we can prove that:

$$(v_m)^n_{i,j,k} \leq v^{n+1}_{i,j,k} \leq (v_M)^n_{i,j,k}\,. \tag{7}$$

Here $l = \{E, W, N, S, F, R\}$ are the mnemonic subscripts for the East, West, North, South, Front and Rear neighboring pixels. As a result, at each iteration the maximum and the minimum of v become closer and no new maxima or minima are created. This characteristic guarantees the stability of the proposed discretization scheme since it prevents the video pixels' intensity from growing in time.

B. Stopping Criterion

As discussed in Sec. II-B, our diffusion procedure progressively smoothes regions whose motion is not coherent with the audio channel activity. Looking at one frame we can observe that the intensity of the edges becomes close to their entourage, but *the same happens across frames*. Thus, the temporal edges in non-relevant regions are iteratively smoothed and the motion which is not related to the soundtrack is reduced. However, if the diffusion process is not stopped it would finally blur the entire signal, eroding also the audio-related parts. In this paper we define a stopping criterion for the audio-visual diffusion process which is intuitive and has a low computational cost.

Let $\mathcal{L}$ be a subset of the video domain Ω: $\mathcal{L} \subset \Omega$. Then, the *amount of motion* M in the video subset $\mathcal{L}$ at iteration n is defined as

$$M^n_{\mathcal{L}} := \sum_{\{i,j,k\}\in\mathcal{L}} |\delta^*_t v^n_{i,j,k}|\,, \tag{8}$$

where $|\delta^*_t v^n_{i,j,k}|$ is the absolute value of the temporal derivative approximation $\delta^*_t v$ at pixel coordinates $\{i, j, k\}$:

$$\delta^*_t v_{i,j,k} = \frac{v_{i,j,k+1} - v_{i,j,k-1}}{2\Delta t}\,. \tag{9}$$

As shown in Fig. 4 [left], the *amount of motion* in the entire video domain ($M^n_\Omega := M^n$) decreases through iterations because at each point the absolute value of the discrete temporal derivative $|\delta^*_t v|$ is bounded by

$$|\delta^*_t v^n_{i,j,k}| \leq \frac{(v_M)^n_{i,j,k} - (v_m)^n_{i,j,k}}{2\Delta t}\,, \tag{10}$$

which is monotonically decreasing (see Eq. (7)). Our method iteratively eliminates the motion in regions that are not related

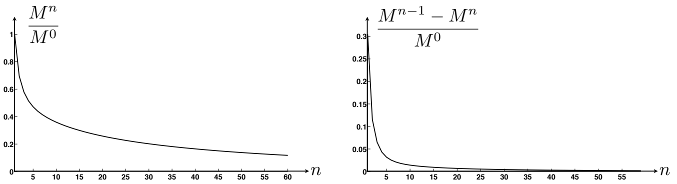


Fig. 4. Typical form of the evolution through iterations of the amount of motion in the video signal [left] and the corresponding motion reduction [right].

to the audio signal and leads to a global reduction of the motion in the video domain. In Fig. 4 [left] only 20% of the original amount of motion M^0 is kept after $n = 40$ iterations. The rest (80%) is considered as non-related to the soundtrack and is iteratively removed. The shape of this curve depends on the parameters choice. Thus, for example a higher $\Delta\tau$ represents a faster decrease in M^n since we converge faster towards the solution. In any case, the decrease on the amount of motion is smaller through iterations, tending towards a relatively stable value.

According to this observation, we define the *motion reduction* ΔM at iteration n as

$$\Delta M^n := \frac{M^{n-1} - M^n}{M^0} \, . \tag{11}$$

This relative value denotes the percentage of the video motion that is eliminated by our algorithm at iteration n. Thus, when the amount of motion does not decrease significantly $\Delta M^{n_{stop}} < \epsilon$ we stop the diffusion process since we consider that most of the information in regions that are not related to the soundtrack has already been eliminated and we are close to the resultant motion map. Fig. 4 [right] represents a typical shape of the evolution of ΔM through iterations. Here 30% of the video motion has been removed at the end of iteration 1, while iteration 10 only eliminates the 1% of the original motion. In this work, the value of ϵ has been fixed to $\epsilon = 0.005$. Here we consider that a reduction of 0.5% is not worth the computation of another iteration since it does not change the motion map in a significant way.

IV. Equalization of Audio and Video Features

Some considerations should be taken into account regarding the audio and video features that we use in Eq. (5) to estimate the audio-video synchrony. As explained in Sec. II-B, the audio feature $a(t)$ represents the energy in the audio channel and the video feature $\partial_t v$ corresponds to the motion in the video signal. However both features have been processed to improve the performance of the proposed method. Thus, the audio feature $a(t)$ is an *equalized* audio energy, while the video feature $\partial_t v$ is also an *equalized* video motion, which means that all the "peaks" in each domain have approximately the same magnitude. This ensures that our approach will give the same opportunities to all the significant motion and sounds instead of prevailing only the most intense video motion occurring exactly at the same time as the louder sound. As a result, the movements that are related to the soundtrack can be effectively preserved even if they are significantly smaller than some distracting motion in the scene.

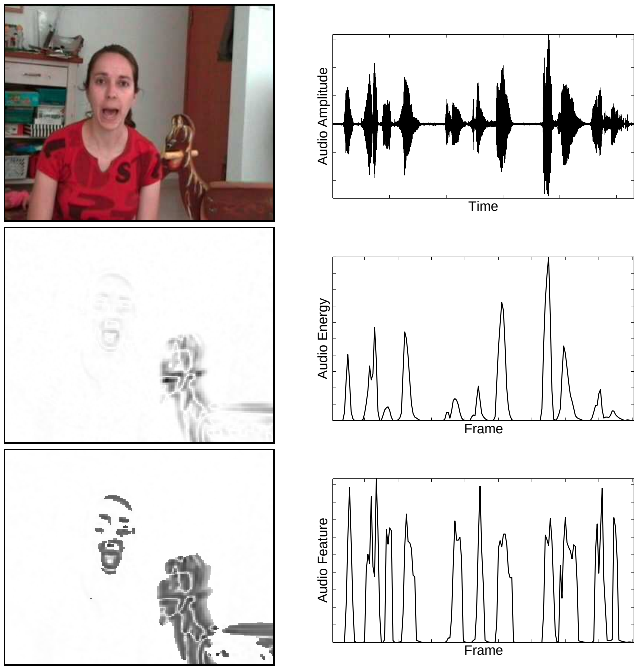


Fig. 5. Proposed features [bottom] corresponding to the audio and video signals in the top row. Right column shows [from top to bottom] the original audio signal, its energy and the equalized energy $a(t)$ at the same temporal resolution than the video signal. Left column depicts one video frame, the motion in this frame (magnitude of the pixels' temporal variation) and the corresponding equalized motion, that is $\partial_t v(x, y, t)$ for a fixed time t. White regions represent static pixels.

The *equalization* in audio and video domains is performed independently but following the same procedure. First, we convolve the original signal with two Gaussians of different variances. We use 3D Gaussians in the case of the video motion and 1D Gaussians for the audio energy. Then, the equalized features are the result of dividing the signals after convolution with the thinner (dividend) and thicker (divisor) Gaussians. Thus, each peak in audio energy and video motion is compared to the energy/motion in the region around it and audio and video features become relative measures.

Some examples of the original and the equalized audio and video features can be observed in Fig. 5. The audio feature [bottom right] has approximately the same magnitude for all significant sounds recorded with the microphone even if originally they had very different energy. Regarding the video signal, the strong motion corresponding to a rocking horse and the mouth movements which are hardly visible in Fig. 5 [center left] are also represented by a similar magnitude in the video feature [bottom left].

Other features for audio and video signals could also be used. For example in the audio case we could use a smoothed version of a binary audio activity detector, the acoustic energy in an important audio sub-band or a measure of the audio nonstationarity. More complex features could also be considered in the video case, but their computation should have a low complexity. Notice that the video feature needs to be recomputed through the diffusion procedure: the audio-visual synchrony s_σ at diffusion time τ depends on $\partial_t v$ at time τ and thus on the evolving video volume itself. Thus, the use of optical flow instead of the temporal derivative of the video signal (computed by means of finite differences) would represent an important increase in terms of computational cost. In this work we prefer to have a very basic but *fast* estimate of the possible locations of the sound sources and use the nonlinear diffusion procedure to ensure spatio-temporal consistency. As a final remark, we stress that the features should not be very selective since audio and video channels are never exactly synchronous.

Results obtained when using different features are shown in the experiments section.

V. Efficiency Measure

We propose a measure to quantify the efficiency of the proposed method in removing the video information that is not related to the sounds in the audio channel.

First, we define an audio-visual region of interest (ROI) as the subset of pixels in the video domain whose motion is related to the soundtrack and the complementary region ($\overline{\mathrm{ROI}}$) as the rest of pixels in the video domain: $\mathrm{ROI} \cup \overline{\mathrm{ROI}} = \Omega$. Then, the *audio-visual diffusion ratio* α at iteration n can be defined as

$$\alpha^n = \left[\frac{\frac{M^0_{\mathrm{ROI}}}{M^n_{\mathrm{ROI}}}}{\frac{M^0_{\overline{\mathrm{ROI}}}}{M^n_{\overline{\mathrm{ROI}}}}} \right]_{a^{ON}} , \qquad (12)$$

where the value $M^0_{\mathrm{ROI}}/M^n_{\mathrm{ROI}}$ is the ratio between the amount of motion *inside* the region of interest at iterations 0 (original motion) and n, and $M^0_{\overline{\mathrm{ROI}}}/M^n_{\overline{\mathrm{ROI}}}$ is the same ratio computed *outside* this region of interest. Here $[\cdot]_{a^{ON}}$ indicates that only the frames where the audio channel is active (a^{ON}) are used in the computation of this ratio. In this work we consider the audio channel to be active when sounds are captured by the microphone and thus the normalized audio feature is large enough: $a(t) > 0.1$ with $a(t) \in [0, 1]$. Thus, the *audio-visual diffusion ratio* α is a relative measure that assesses the ability to attenuate the motion in parts of the video signal that are not related to the soundtrack by comparing it to the diffusion experienced in the audio-visual region of interest, *when sounds are present in the audio channel*. $\alpha > 1$ when our method favors regions associated to the soundtrack, $\alpha = 1$ if the video motion is equally eliminated inside and outside the ROI, and $\alpha < 1$ when the diffusion affects more the ROI than the rest of the video signal in non-silent periods. Notice that obtaining $\alpha > 1$ is an extremely challenging task, especially in sequences where the audio-related motion is less intense than the distracting motion.

VI. Experiments

The evaluation has been performed in sequences of different nature presenting strong auditive and/or visual distractors. All the sequences are composed of two moving objects, and only one of them is related to the soundtrack. The purpose of this configuration is to allow a quantitative comparison between the strength with which the diffusion process affects the audio-related region and the distracting moving object by means of the efficiency measure α.

MovieA and MovieB (Fig. 6) are taken from the state-of-the-art source localization work presented by Kidron et al. in [17]. In MovieA the audio signal is generated by a hand playing a guitar and then a synthesizer, while in MovieB we can see a person speaking and the audio signal is corrupted by the voice of another person. A strong periodic visual distraction is introduced by means of a rocking wooden horse. Both video sequences are sampled at 25 frames/sec at resolution of 576×720 pixels and the audio at 44.1 kHz. For its analysis, the video signal has been resized to 144×180 pixels. Each sequence is 10 seconds long approximately.

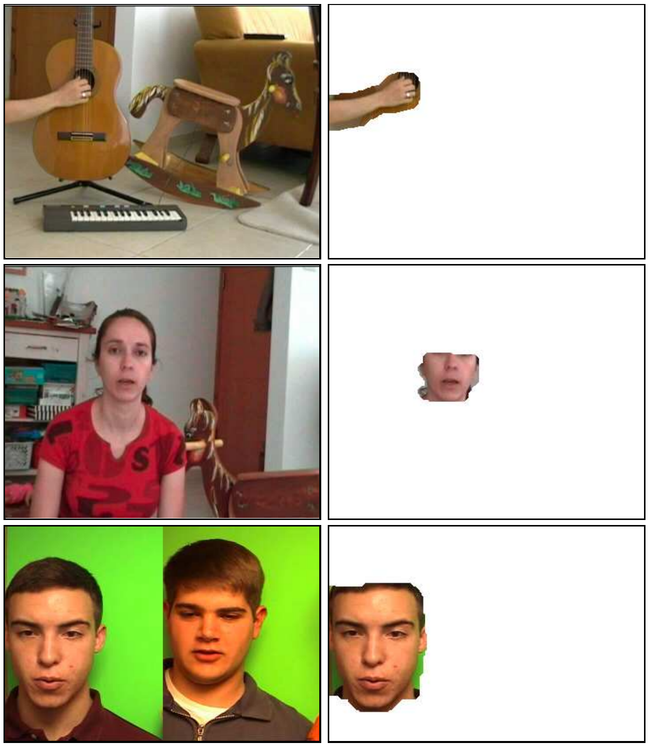


Fig. 6. From top to bottom: frames belonging to MovieA, MovieB and MovieC [left] and corresponding regions of interest (ROI) [right] used to evaluate quantitatively the proposed method. White regions in the right column depict parts of the image not related to the soundtrack ($\overline{\mathrm{ROI}}$).

A third sequence, MovieC, is synthesized using clips g01 and g08 from the *groups* partition of CUAVE database [39]. The video part corresponds to two persons uttering the same numbers in front of a camera but we only keep the audio corresponding to the left person in Fig. 6 [bottom left]. The resulting sequence is thus composed of one person uttering numbers and another one mouthing the same numbers. Thus, in this scene we have again one object (person) contributing to the soundtrack and one strong audio-visual distractor. In this case the motion generated by the distractor (silent person) and the audio-related object are very similar. The video part of MovieC is sampled at 29.97 frames/sec with a resolution of 480×720 pixels, while the audio part is sampled at 44 kHz. For its analysis, the video signal has been resized to 120×176 pixels. This sequence is around 6 seconds long.

This section is organized as follows. Sec. VI-A provides a qualitative analysis of the resulting signals after the proposed nonlinear diffusion procedure. In Sec. VI-B we present a quantitative evaluation of the performance of our method. Finally, Sec. VI-C compares the results when using different

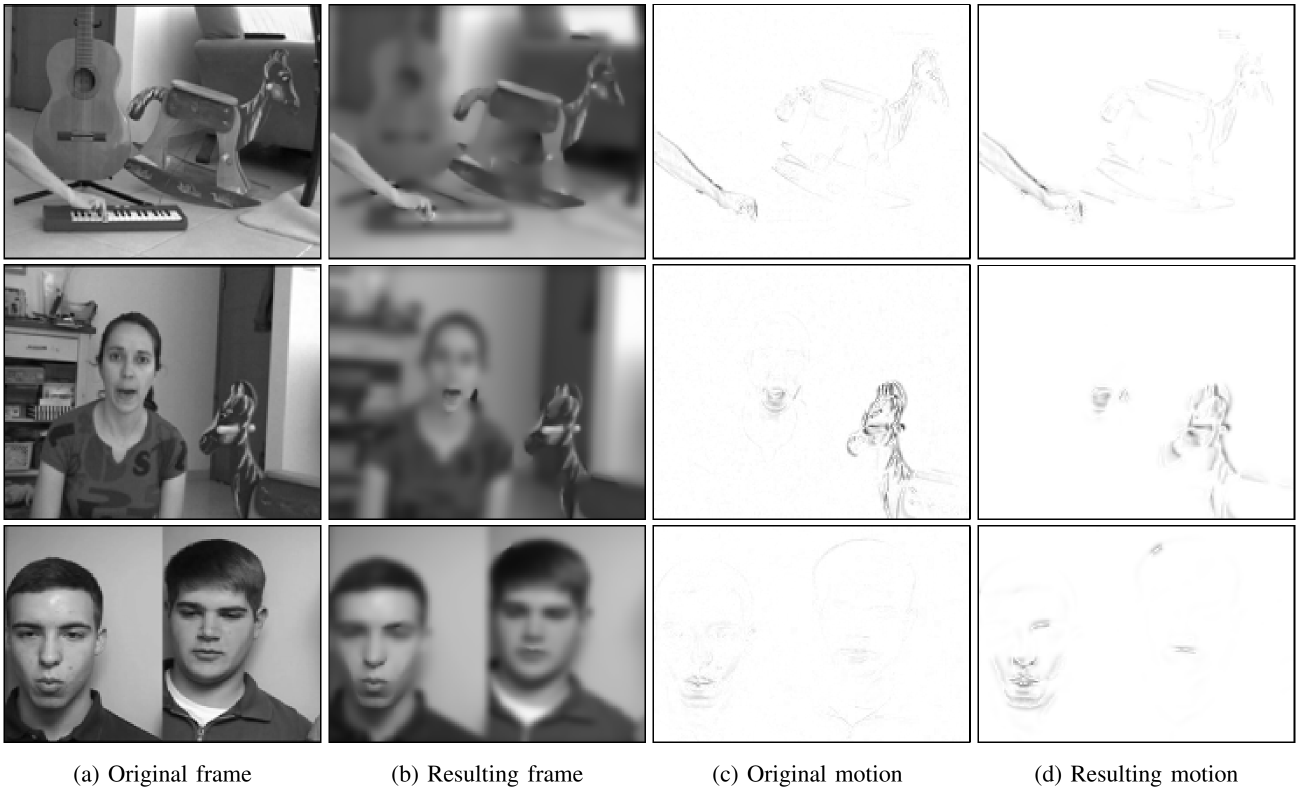


Fig. 7. Results obtained when applying our method to MovieA, MovieB and MovieC with $K = 0.1$. The diffusion process has been automatically stopped after $n_{stop} = 26, 25, 11$ iterations respectively according to the stopping criterion in Sec. III-B.

audio and video features.

We use the same parameters in all experiments. We fix $\sigma = 1$ to avoid artifacts due to noise and ensure spatio-temporal coherence. The parameter that controls the diffusion speed is fixed to $\Delta\tau = 0.15$ since we need $\Delta\tau \in [0, 1/6]$ to satisfy the maximum and minimum principle in Eq. (7). The *audio-visual synchrony* is normalized: $s_\sigma \in [0, 1]$. Different values of K ranging between 0.05 and 0.15 are used for comparative purposes in Sec. VI-B. However, the rest of experiments in this section are performed with $K = 0.1$.

The computational complexity of one iteration of our method is $\mathcal{O}(N \log N)$, where N is the number of pixels in the video volume. The number of required iterations n_{stop} is determined as explained in Sec. III-B.

A. Qualitative Analysis

Results obtained when analyzing MovieA, MovieB and MovieC with the proposed method are shown in Fig. 7. The original frames of those sequences in (a) present a lot of irrelevant background details such as a carpet or small objects in the shelves that completely disappear or become blurred in the resulting frames in (b). Even if the rocking horse is moving continuously, its silhouette is blurred and most of its details disappear equally. In contrast, the focus is preserved in regions related to the soundtrack, i.e. the hand in MovieA, the girl's mouth in MovieB and the left speaker's mouth in MovieC. By comparing columns (c) and (d) it is possible to observe that in all cases the motion is better preserved in the audio-related video regions, even though some situations are really challenging because the distracting motion is much more intense.

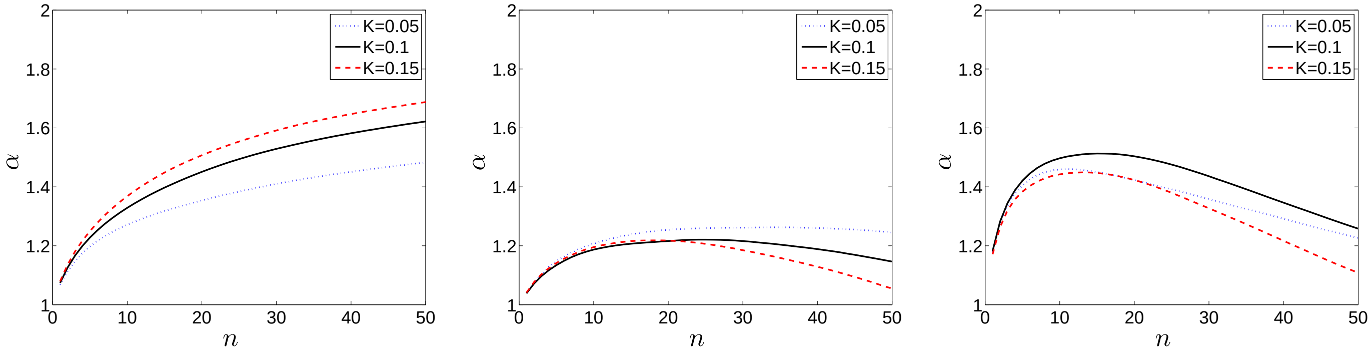
In MovieB the audio signal is corrupted by a second voice. However, the audio feature is not affected by the person speaking out of the field of view, since the energy of this second voice is significantly smaller than the energy of the girl's voice. As a result, the background sounds do not affect significantly the result and the video signal is focused on the girl's mouth only when she is speaking.

Videos showing the test sequences and the corresponding video signals after applying our method are available online at *http://lts2www.epfl.ch/people/llagostera/*.

These experiments illustrate also the limitations of our approach. In fact, when the analyzed sequence contains a distracting motion which is synchronized with the soundtrack, our algorithm is not able to remove it. An example can be found when the two persons in MovieC utter a word exactly at the same time. In this case, the focus is kept in the mouths of both persons because they could *both* be the sound source, i.e. both movements are coherent with the sound. In fact, we could be hearing two words, one uttered by each person, and the audio feature would not change. Some a priori knowledge about the frequency characteristics of their voices might help in discarding the silent person. However, here we want to keep our method general and we only use the assumption of synchrony between audio and video channels.

B. Quantitative Analysis

This section evaluates the efficiency of the proposed non-linear diffusion approach in prevailing the video information



(a) MovieA (b) MovieB (c) MovieC

Fig. 8. Evolution through iterations of the *audio-visual diffusion ratio* α for different values of K.

that is useful in audio-visual analysis. For this purpose we use the audio-visual diffusion ratio α defined in Sec. V, which compares the amount of video motion removed *inside* and *outside* some region of interest (ROI) when sounds are present. In this work, the regions of interest for audio-visual analysis are defined as spatio-temporal regions in the video signal whose motion generates the sounds captured with the microphone. Fig. 6 shows a frame belonging to each test sequence and the corresponding ROI in this frame. From top to bottom, the ROI in MovieA corresponds to the hand that plays the guitar and the piano, in MovieB it is defined as the speaker's mouth region, and it is the speaker's face in MovieC. The depicted ROIs have been manually defined using a 3D video segmentation interface.

Fig. 8 shows the audio-visual diffusion ratio α that we obtain when applying our method to MovieA, MovieB and MovieC with different values for the parameter K in Eq. (6). As expected, in all sequences and for K ranging between 0.05 and 0.015, we obtain satisfactory values for the audio-visual diffusion ratio ($\alpha > 1$). This result proves that the video motion is prevailed more efficiently inside the ROIs when the audio channel is active.

However, there is not an optimal value for K that provides the best performance in all situations. The higher is K the higher is the diffusion coefficient D (see Eq. (6)). As a result, the diffusion process affects the video volume with more strength and the motion in the signal is reduced *faster*. Thus $K = 0.15$ leads to a good performance in MovieA and to a faster removal of the distracting motion. In contrast, high values for K can result in the elimination of information in regions that are associated to the soundtrack if the initial motion in these regions has a low magnitude. In MovieB for example, when $K = 0.15$ the diffusion affects most moving regions almost from the beginning and some audio-related motion in the speaker's mouth is eliminated. A good compromise can be obtained by fixing $K = 0.1$. As shown in Fig. 8, a high audio-visual diffusion ratio α is reached faster when the audio-related video motion is not very small (MovieA and MovieC) and the results when the distracting motion is dominant are also satisfactory (MovieB).

Table I depicts the results obtained for the three analyzed sequences when using the stopping criterion defined in Sec. III-B. First of all, notice that the stopping time determined for MovieB and MovieC leads to values of α that are close to the maximum of curves in Fig. 8. Even if the curve corresponding to MovieA does not present a maximum for a small number of iterations, the diffusion process is stopped when the value of α is high. As discussed before, increasing the number of iterations is not advisable in this case since it increases the computational cost without changing significantly the motion map. Finally, the lowest K ($K = 0.05$) leads to the smallest number of iterations in all cases, since the amount of motion in the video signal decreases slowly. In this case, many motion concentrations are considered as possibly related to the soundtrack, it takes time to discard them and the motion in the video volume evolves so slowly that after some iterations the motion map seems already stuck.

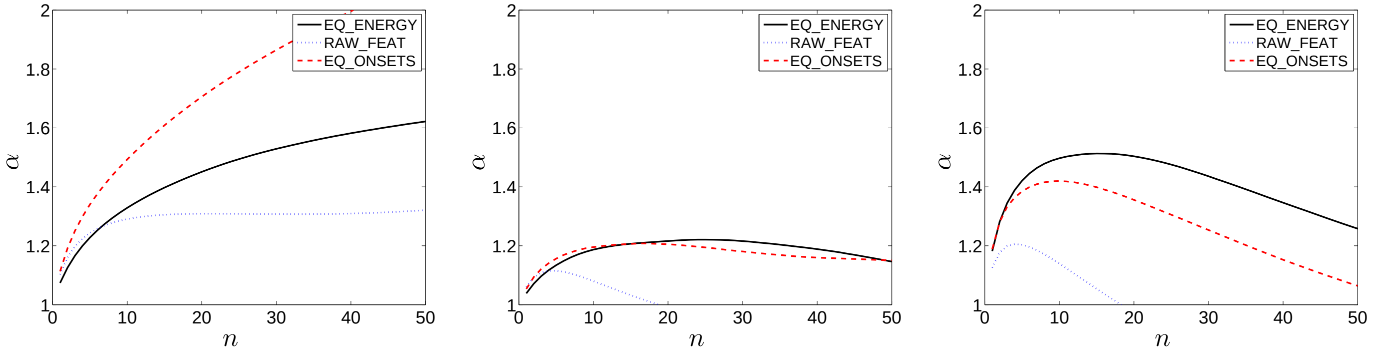
	MovieA	MovieB	MovieC
$K = 0.05$	1.33 (17)	1.26 (23)	1.44 (7)
$K = 0.1$	1.50 (26)	1.22 (25)	1.50 (11)
$K = 0.15$	1.59 (30)	1.20 (26)	1.45 (15)

TABLE I
RESULTING AUDIO-VISUAL DIFFUSION RATIO α FOR DIFFERENT VALUES OF K. THE NUMBER OF ITERATIONS THAT ARE REQUIRED ACCORDING TO THE STOPPING CRITERION ARE SHOWN IN PARENTHESIS.

C. Feature Selection

Finally, we compare the performance of the proposed audio-visual diffusion procedure when using different features (see Table II) for the computation of the audio-video synchrony measure s_σ in Eq. (5). The purpose of this section is to demonstrate the effectivity of the equalization step in Sec. IV and to compare the performance of the audio energy to another feature that is commonly adopted in audio-visual fusion, i.e. the onsets in the audio channel.

Audio onsets represent a measure of the nonstationarity in the audio channel and they are used in other audio-visual fusion methods [18]. Since there are multiple examples of stationary sounds that do not have any motion associated (e.g. a car engine), in some situations audio onsets might perform better than the audio energy in assessing the synchrony between audio and video channels. In our case, onsets are obtained by computing the time derivative of the audio energy



(a) MovieA (b) MovieB (c) MovieC

Fig. 10. Evolution through iterations of the *audio-visual diffusion ratio* for the three different combinations of audio and video features in Table II.

	AUDIO FEATURE	VIDEO FEATURE
EQ_ENERGY	Equalized energy	Equalized motion
RAW_FEAT	Energy	Motion
EQ_ONSETS	Equalized onsets	Equalized motion

TABLE II
THREE TESTED COMBINATIONS OF AUDIO AND VIDEO FEATURES. OUR METHOD USES EQ_ENERGY.

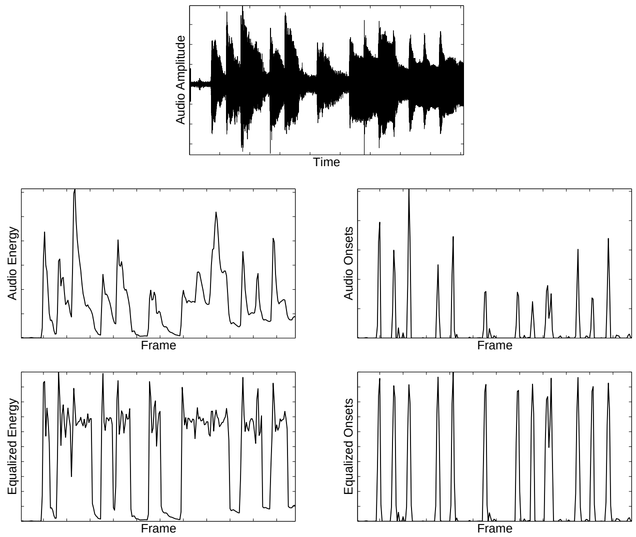


Fig. 9. Soundtrack belonging to MovieA [top] and corresponding *equalized* audio energy [bottom left] and onsets [bottom right].

as explained in [40]. Fig. 9 [right] shows an example of audio onsets before and after equalization.

Fig. 10 shows the resulting audio-visual diffusion ratio α for the analyzed sequences when the features are chosen according to the three options in Table II. In all cases RAW_FEAT performs worse than the other two possibilities, leading to $\alpha < 1$ after less than 20 iterations for MovieB and MovieC. This result demonstrates the importance of the equalization process in Sec. IV, which ensures equal opportunities to all significant sounds and motion. Regarding EQ_ONSETS, its performance is superior than EQ_ENERGY in MovieA, similar in MovieB and worse in MovieC. MovieA contains piano sounds, each of them composed by an onset followed by a decay in the acoustic energy (see Fig. 9 [top]). The video motion in this case is synchronized with the onsets and not with the decay period. The equalized onsets [bottom right] capture only the time instants in which the keys are pressed and thus the distracting motion can be effectively attenuated. In contrast, the value of the equalized audio energy [bottom left] is high during periods in which there is no motion correlated to the soundtrack but the distracting motion is still present. While the onsets seem more adequate than the audio energy when the soundtrack contains stationary sounds, the equalized audio energy leads to a better performance in sequences containing speakers (MovieB, MovieC).

VII. APPLICATION: UNSUPERVISED EXTRACTION OF AUDIO-RELATED VIDEO REGIONS

The proposed audio-visual diffusion procedure erodes video regions presenting a low coherence with the audio signal and automatically highlights the possible sound sources. Thus, an intuitive application of this diffusion procedure can be the unsupervised extraction of audio-related video regions. The algorithm that we introduce in this section is very simple, and its purpose is to illustrate the capabilities of our approach. Here we propose first to determine possible regions of interest by comparing the motion before and after the audio-visual diffusion process and then use this knowledge as a starting point for a standard segmentation procedure using graph cuts. The extracted region contains thus the video parts whose motion is highly synchronous to the soundtrack that are identified by the proposed method.

For this purpose we define the *audio-visual coherence* $c(\mathbf{x})$ at pixel location $\mathbf{x}$ as

$$c(\mathbf{x}) = \begin{cases} \frac{\partial_t v(\mathbf{x},\tau_{stop})}{\partial_t v(\mathbf{x},0)} & \text{if} \quad \partial_t v(\mathbf{x},0) > \xi \\ \frac{\partial_t v(\mathbf{x},\tau_{stop})}{\max_{\mathbf{x}} \partial_t v(\mathbf{x},0)} & \text{otherwise} \end{cases} \qquad (13)$$

where $\partial_t v(\mathbf{x}, \tau_{stop})$ is the temporal derivative of the resulting video signal after n_{stop} iterations of the proposed nonlinear diffusion procedure ($\tau_{stop} = n_{stop}\Delta\tau$) and the constant ξ makes the audio-visual coherence $c(\mathbf{x})$ close to zero in static pixels (we can fix $\xi = 10^{-1}$ for example). The higher is the audio-visual coherence $c(\mathbf{x})$ the higher is the probability for the video pixel at location $\mathbf{x}$ to be part of an audio-related

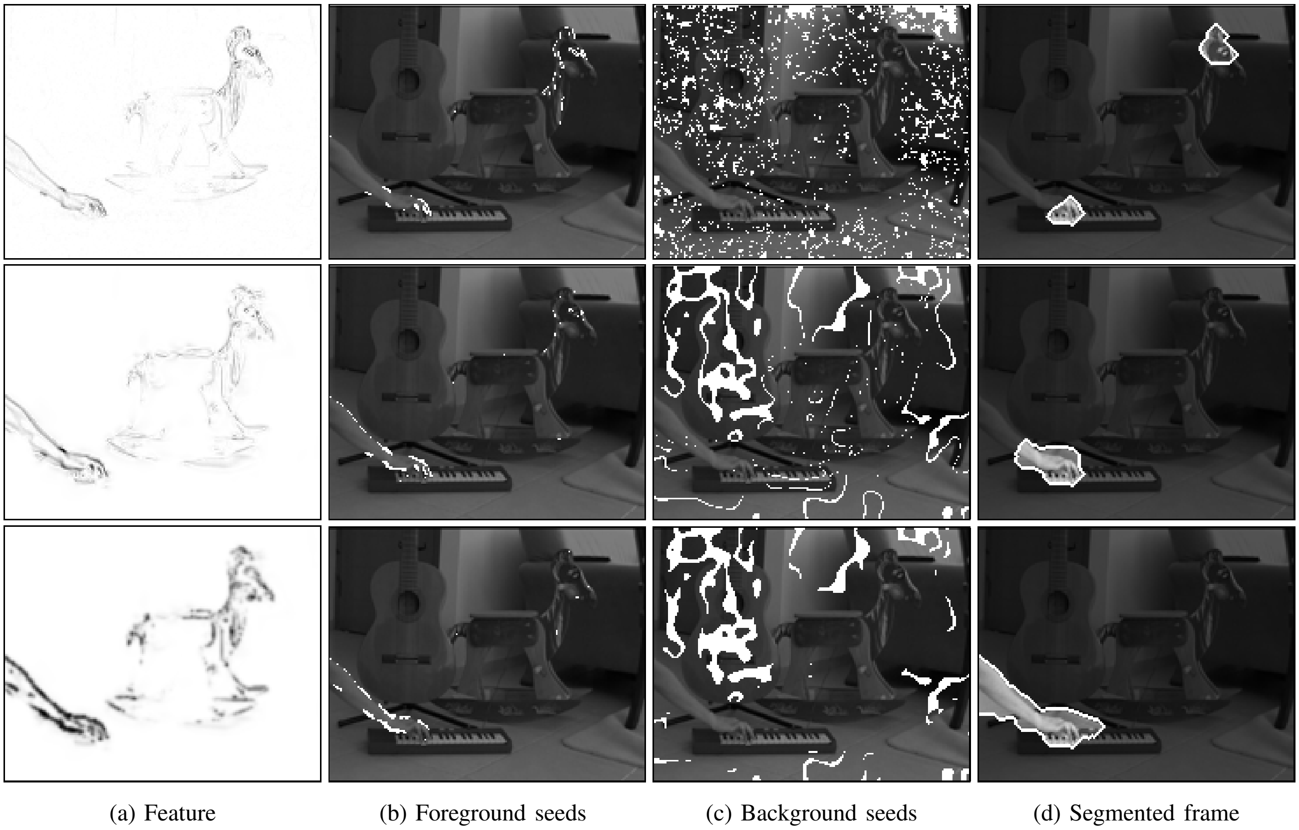


Fig. 11. Extracted audio-related video regions (d) for a frame belonging to MovieA when choosing the segmentation seeds according to the features in (a): original motion [top], resulting motion [middle], and audio-visual coherence $c(\mathbf{x})$ [bottom]. White pixels in (b)-(c) indicate the automatically labeled segmentation seeds. The extracted regions in (d) are delimited by a white line and they are depicted in a brighter grayscale than the background.

video region, since its motion is well preserved through the diffusion process.

The pixels with highest audio-visual coherence $c(\mathbf{x})$ are then labelled as belonging to the audio-related video region (foreground), and the pixels presenting the lowest $c(\mathbf{x})$ are used as background *seeds* (i.e. initial labels). Only a small number of pixels are labelled in this step. Thus, we estimate that the points whose motion is better preserved through the diffusion process are likely to compose the audio-related video region. Once these pixels are automatically labelled, a standard binary segmentation using graph cuts [41] is applied to extract the whole audio-related video region (and label all remaining pixels).

Fig. 11 shows an example of applying this procedure to a frame belonging to MovieA. It compares the extracted regions (d) when we label the pixels according to the original video motion [top], the resulting video motion after the diffusion procedure [middle] and, as proposed in this section, the *audio-visual coherence* $c(\mathbf{x})$ [bottom]. In this case, a 0.5% and a 10% of pixels are automatically labelled as foreground and background in (b)-(c) respectively. Notice that we are much more selective when choosing the foreground seeds, since we want to be sure of labeling only the right pixels. The background seeds in Fig. 11 (c) are well distributed across the frame for the three features that we consider. Regarding the foreground seeds in (b), while they are equally distributed between the hand (audio-related video region) and the horse's head (distracting moving object) according to the initial motion [top], when using the resulting motion [middle] most of them are in the correct location. Finally, the feature that we propose, i.e. the *audio-visual coherence* [bottom], leads to the smallest number of errors on the seed choice, i.e. only a few seeds are located over the rocking horse. Since the extracted region in (d) is determined by the seeds, the *audio-visual coherence* provides more accurate results than the other two features. The extracted audio-related video region in this case [bottom] is very similar to the ROI that was manually defined for the quantitative evaluation (see Fig. 6).

The interested reader can find in [42] a more elaborated approach for the unsupervised extraction of audio-related regions which is also based on the proposed audio-visual diffusion procedure.

VIII. Discussion

We have proposed a novel nonlinear video diffusion approach which is controlled by the fusion of information in audio and video channels. Our method integrates the main assumption in the audio-visual domain in the definition of the diffusion coefficient, which depends on an estimate of the synchrony between video motion and audio energy. As a result, video parts that are related to the synchronously recorded soundtrack are automatically highlighted while information which is not useful for audio-visual applications is progressively reduced.

Several tests have been performed in challenging real-world sequences. Quantitative results show that our approach is effective in prevailing audio-related video regions over other moving objects. However, our method is unable to distinguish between two regions whose motion is coherent with a sound. When two persons mouth a word at the same time for example, both mouth regions are highlighted independently of which voice we hear. We do not want to introduce any additional knowledge about the sources' characteristics because our goal is to keep this method as general as possible. We believe that this approach can be efficiently used as a preprocessing step for other methods in this domain, since it is able to remove misleading information in applications such as sound source localization.

References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[2] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds. Lawrence Erlbaum Associates, 1987, pp. 3–51.

[3] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading." *Nature*, vol. 381, no. 6577, pp. 66–68, 1996.

[4] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition," *IEEE Trans. on Multimedia*, vol. 7, no. 3, pp. 495–506, 2005.

[5] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[6] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[7] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2002.

[8] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Commununication*, vol. 44, no. 1-4, pp. 113–125, 2004.

[9] R. Dansereau, "Co-channel audiovisual speech separation using spectral matching constraints," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2004.

[10] S. Rajaram, A. V. Nefian, and T. Huang, "Bayesian separation of audio-visual speech sources," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2004.

[11] P. Pérez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proc. of the IEEE*, vol. 92, no. 3, pp. 495–513, 2004.

[12] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2000.

[13] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proc. of Int. Conf. Image and video retrieval (CIVR)*, 2003.

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

[15] P. Smaragdis and M. Casey, "Audio/visual independent components," *Proc. of Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003.

[16] M. Siracusa and J. Fisher, "Dynamic dependency tests: Analysis and applications to multi-modal data association," in *Int. Conf. Artificial Intelligence and Statistics (AIStats)*, 2007.

[17] E. Kidron, Y. Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. on Signal Processing*, vol. 55, no. 4, pp. 1390–1404, 2007.

[18] Z. Barzelay and Y. Y. Schechner, "Harmony in motion," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007.

[19] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind Audio-Visual Source Separation based on Sparse Redundant Representations," *IEEE Trans. on Multimedia*, vol. 12, no. 5, pp. 358–371, 2010.

[20] C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann, "Nonnegative cca for audiovisual source separation," in *IEEE Workshop on Machine Learning for Signal Processing*, 2007.

[21] G. Chetty and M. Wagner, "Audio visual speaker verification based on hybrid fusion of cross modal features," in *Pattern Recognition and Machine Intelligence (PReMI)*, 2007.

[22] C. Saraceno and R. Leonardi, "Indexing audiovisual databases through joint audio and video processing," *Int. Journal of Imaging Systems and Technology*, vol. 9, no. 5, pp. 320–331, 1999.

[23] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer, "Audiovisual person tracking with a mobile robot," in *Proc. of Int. Conf. Intelligent Autonomous Systems*, 2004.

[24] J. Hershey and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 1999.

[25] G. Monaci, O. Divorra, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Processing*, vol. 86, no. 12, pp. 3534–3548, 2006.

[26] P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Trans. on Multimedia*, vol. 10, no. 1, pp. 63 –73, 2008.

[27] R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio correlation," in *Proc of. IEEE Int. Conf. Multimedia and Expo (ICME)*, 2000.

[28] 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.

[29] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval, "Learning Multi-Modal Dictionaries," *IEEE Trans. on Image Processing*, vol. 16, no. 9, pp. 2272–2283, 2007.

[30] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[31] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*, ser. Applied Mathematical Sciences. Springer, 2006, vol. 147.

[32] F. Catte, P. Lions, J. Morel, and T. Coll, "Image selective smoothing and edge detection by nonlinear diffusion," *SIAM Journal on Numerical Analysis*, vol. 29, no. 1, pp. 182–193, 1992.

[33] J. Weickert, *Anisotropic Diffusion in Image Processing*. Stuttgart, Germany: Teubner, 1998.

[34] P. Mrázek, "Nonlinear diffusion for image filtering and monotonicity enhancement," PhD Thesis, Czech Technical University, Prague, Czech Republic, 2001.

[35] A. Llagostera Casanovas and P. Vandergheynst, "Audio-based nonlinear video diffusion," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2010.

[36] L. Nirenberg, "A strong maximum principle for parabolic equations," *Communications on Pure and Applied Mathematics*, vol. 6, pp. 167–177, 1953.

[37] G. Monaci and P. Vandergheynst, "Audiovisual gestalts," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2006.

[38] J.-S. Lee and T. Ebrahimi, "Efficient video coding in H.264/AVC by using audio-visual information," in *Proc. of IEEE Int. Workshop on Multimedia Signal Processing (MMSP'09)*, 2009.

[39] 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*, vol. 2002, no. 11, p. 1189, Nov. 2002.

[40] J. P. Bello, L. Daudet, S. A. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals." *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[41] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proc. of IEEE Int. Conf. Computer Vision (ICCV)*, 2001.

[42] A. Llagostera Casanovas and P. Vandergheynst, "Unsupervised Extraction of Audio-Visual Objects," in *Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2011.