AUDIO-BASED NONLINEAR VIDEO DIFFUSION

Anna Llagostera Casanovas and Pierre Vandergheynst

Signal Processing Institute (LTS2), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

We propose a novel non-linear video diffusion approach which is able to focus on parts of a video sequence that are relevant for applications in audio-visual analysis. The diffusion process is controlled by a diffusion coefficient based on an estimate of the synchrony between video motion and audio energy at each point of the video volume. Thus, regions whose motion is not coherent with the soundtrack are iteratively smoothed. The discretization of the proposed continuous diffusion formulation is carefully studied and its stability demonstrated. Our approach is tested in challenging situations involving sequence degradation and distracting video motion. Results show that in all cases our method is able to keep the focus of attention on the sound sources.

Index Terms— Audio-visual processing, linear/nonlinear diffusion, finite difference methods

1. INTRODUCTION

Natural scenes are very rich and it is difficult to understand them properly without using information coming from different sensory modalities. In order to emulate human behavior, scientists have been trying to combine different research domains. In particular, audio-visual analysis aims to put together information coming from audio and video channels. Thus, it is possible to use the video information to improve results in the audio domain for applications such as speech recognition, speech enhancement [4] and sound source separation [12]. Other methods try to assess coherence between both modalities to track or locate [5, 9] sound sources in the video signal. Some latter approaches go further and try to separate a scene into audio-visual structures, each of them composed by a visual part and the associated soundtrack [2, 6].

The challenge in audio-visual analysis lies in efficiently combining the 3-D video signal recorded with a video-camera and the corresponding 1-D audio signal coming from *one microphone*. Fusion of audio and video modalities is still an open problem. Most of the approaches in audio-visual analysis first define features for each modality such as audio energy [5, 11] or cepstrum coefficients [9] for the audio, and pixel intensities [11] or temporal variations [5, 9] for the video. Then, they use these representations in a fusion step, which try to assess the synchrony between both modalities using canonical correlation analysis [5] or joint audio-visual probabilities for the features [11]. All those methods are based on pixel behavior, which makes them vulnerable to visual noise and does not ensure video spatial coherence. Other approaches propose to decompose each modality [6] or both modalities at the same time [7] over redundant dictionaries of signals. That makes the fusion step cheap and intuitive since we deal now with a small number of video *structures*. However, this decomposition is time consuming.

This work was supported by the Swiss NFS through the IM.2 National Center of Competence for Research and grant number 200021-117884.

Most applications in audio-visual analysis only use video parts that are coherent with the soundtrack, i.e. speech recognition only needs the speaker's mouth, and sound source localization is based on regions moving coherently with sounds. Even if the remaining video information is not needed, identifying a mouth or discriminating relevant motion from distracting motion involves a significant amount of computational cost. The aim of this research work is to simplify audio-visual sequences by eliminating most of this non-relevant video information through a cheap and fast procedure.

The method we present here is able to homogeneously diffuse parts of the video signal whose motion is not coherent with a synchronously recorded audio track while keeping the rest. The diffusion process is controlled by a diffusion coefficient that is a function of the synchrony between audio energy and video motion at each point of the video signal. An accurate discretization scheme is proposed in order to ensure the diffusion process stability and to prevent the creation of new maxima. Several tests are performed in challenging real-world sequences presenting degraded signals and important distracting video motion. Results show that the proposed approach is able to automatically focus on the sound source region. As a result, the relevant movements are kept and the distracting video motion is eliminated or attenuated.

Sec. 2 recalls the main principles of PDE-based diffusion, while Sec. 3 presents the proposed continuous model for nonlinear audio-based video diffusion. In Sec. 4 the numerical scheme used for the problem discretization is detailed and its properties are discussed. Sec. 5 presents the obtained results when analyzing challenging natural audio-visual sequences. Finally, in Sec. 6 achievements and future research directions are discussed.

2. DIFFUSION BACKGROUND

Let us consider a 3-D video domain $\Omega := (0, b_1) \times (0, b_2) \times (0, b_3)$ 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 :

$$\partial_\tau v = \mathrm{div}(D\nabla v) \quad \text{on} \quad \Omega \times (0, \infty)\,, \tag{1}$$

$$v(\mathbf{x}, 0) = f(\mathbf{x}) \quad \text{on} \quad \Omega\,, \tag{2}$$

$$\langle D\nabla v, \mathbf{n}\rangle = 0 \quad \text{on} \quad \Gamma \times (0, \infty)\,, \tag{3}$$

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

The diffusion equation in (1) belongs to a general class of equations satisfying the *maximum principle* (see the proof in [8]). The

principle states that all the maxima of a solution of equation (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 in our case the diffusion is 0 across the boundary Γ (see equation (3)) the maxima can only belong to the original image (initial condition at $\tau = \tau_0$). In practice, the principle prevents the creation of new local extrema when applying the diffusion process to any function v.

Chronologically, applications in the signal processing domain have evolved from using simple constant values for the diffusion coefficient D until much more complex expressions :

1. *Linear diffusion*: The diffusion coefficient is constant on space and diffusion time : $D(\mathbf{x}, \tau) = c$, where c is a scalar. The solution of the resulting diffusion equation $\partial_\tau v = c\Delta v$ is equivalent to convolving the original video signal $v(\mathbf{x}, 0)$ with a Gaussian of variance $\sigma = \sqrt{2\tau}$ (see [13] for a more detailed explanation).
2. *Scalar-valued nonlinear diffusion*: At each point $\mathbf{x}$ and iteration step τ the diffusion coefficient is represented by a scalar value : $D(\mathbf{x}, \tau) \in \mathbb{R}$, $\forall\, \mathbf{x}, \tau$. In fact, the diffusion coefficient depends on the evolving video signal itself. This model was first proposed by Perona and Malik in [10] and it is commonly applied to edge detection.
3. *Vector-valued nonlinear diffusion*: The diffusion is controlled by a tensor depending on $\mathbf{x}$ and τ: $\mathbf{D}(\mathbf{x}, \tau) \in \mathbb{R}^{3\times 3}$, $\forall\, \mathbf{x}, \tau$. This characteristic gives more freedom to the diffusion process and it can be applied to detection of corners or line-like structures (see Weickert's work in [13]).

3. AUDIO-VISUAL DIFFUSION

In this paper we propose the following scalar-valued diffusion coefficient D :

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

where the function $g(\cdot)$ determines the intensity of the diffusion process given s_σ, which is a regularized measure of synchrony between audio and video channels defined as :

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

Here, G_σ is a 3-D Gaussian of variance σ^2, $a(x, y, t) = a(t)\ \forall x, y$ represents the energy on the audio channel at time t and $\partial_t v$ is the temporal derivative of the video signal. 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. s_σ is high when an important acoustic event matches a relevant pixel motion while its value is close to 0 in the rest. Other possibilities for the audio feature are a smoothed version of a binary activity detector or the acoustic energy in an important audio sub-band. In fact, $a(t)$ should not be very selective since audio and video channels are never exactly synchronous. The convolution with a Gaussian G_σ in expression (5) makes our audio-visual synchrony measure s_σ much more robust to visual and acoustic noise. Furthermore, this procedure has been used by Catté et al. in [3] in order to regularize the nonlinear diffusion problem presented by Perona and Malik in [10], whose formulation is similar to ours.

Let us now discuss the shape of the function $g(\cdot)$ in equation (4). First of all, we want a linear diffusion process to take place in spatio-temporal regions with low audio-visual synchrony s_σ. In addition, the diffusion coefficient D should be close to 0 in points with high synchrony s_σ in order to stop there the diffusion. Finally, the diffusion coefficient D has to be positive to accomplish the maximum principle. An appropriate function $g(\cdot)$ can thus be the non-negative monotonically decreasing function proposed by Perona and Malik in [10] (see Fig. 1) :

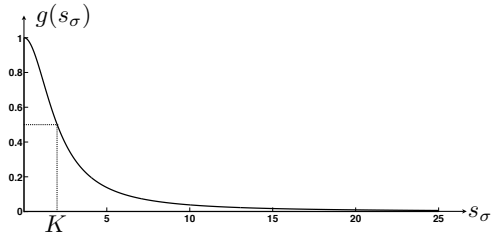


Fig. 1. Shape of the function $g(\cdot)$ in equation (6).

$$g(s_\sigma) = \frac{1}{1 + \left(\frac{s_\sigma}{K}\right)^2}\,. \tag{6}$$

The value of the constant K should be chosen carefully since it acts as a threshold : points where $s_\sigma < K$ are strongly affected by linear diffusion while points where $s_\sigma > K$ are least diffused. The value of K can then be either set by the user or fixed for example to some percentile of the initial distribution of the variable s_σ ($\tau = 0$). Notice that if we plot an histogram of s_σ most of the values are close to 0, since a great majority of the points have either a negligible video motion or a small audio energy.

Let us analyze qualitatively the behavior of the proposed audio-visual diffusion process. 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).

Since *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 linearly diffused. Out of those regions, the diffusion coefficient D becomes smaller and the diffusion process is stopped. Furthermore, the nature of linear diffusion together with the regularization with a Gaussian in equation (5) bring implicitly spatial coherence to our approach by prevailing structures over pixels. Notice that the diffusion coefficient $D \approx 1$ in a single pixel surrounded by pixels with low audio-visual coherence, independently of the synchrony of the pixel itself.

An important consequence of applying the proposed diffusion procedure to an audio-visual sequence is the iterative elimination of the video motion which is not related to the soundtrack. Indeed, spatio-temporal edges situated in regions whose motion is not coherent with the audio channel activity are progressively smoothed since those regions are affected by linear 3-D diffusion. As a result, the variation of the pixels intensity across frames decreases and the video motion is reduced. In fact, by observing the resulting video motion after some iterations of the proposed method we can discover where our algorithm places its attention, that is the possible location of the sound sources in the image.

4. DISCRETIZATION

The proposed discretization scheme has been studied and advised by Aubert and Kornprobst in [1]. The continuous diffusion equation in

(1) can be rewritten as :

$$\partial_\tau v = \partial_x(D\partial_x v) + \partial_y(D\partial_y v) + \partial_t(D\partial_t v)\,. \tag{7}$$

The left part has been discretized following a *forward* finite difference scheme as commonly done in literature [1] :

$$\partial_\tau v|_{ijk}^n \approx \delta_\tau^+ v_{i,j,k}^n := \frac{v_{i,j,k}^{n+1} - v_{i,j,k}^n}{\Delta\tau}\,, \tag{8}$$

where $v_{i,j,k}^n$ is 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 ($\Delta x = \Delta y = \Delta t = h = 1$), while $\Delta\tau$ is the grid spacing used for the diffusion time discretization and controls the diffusion speed. Then, the derivative discretizations in the x, y and t directions are defined as :

$$\partial_x v|_{ijk} \approx \delta_x^* v_{i,j,k} := \frac{v_{i+\frac{1}{2},j,k} - v_{i-\frac{1}{2},j,k}}{\Delta x}\,, \tag{9}$$

$$\partial_y v|_{ijk} \approx \delta_y^* v_{i,j,k} := \frac{v_{i,j+\frac{1}{2},k} - v_{i,j-\frac{1}{2},k}}{\Delta y}\,, \tag{10}$$

$$\partial_t v|_{ijk} \approx \delta_t^* v_{i,j,k} := \frac{v_{i,j,k+\frac{1}{2}} - v_{i,j,k-\frac{1}{2}}}{\Delta t}\,. \tag{11}$$

The values of v at location $((i \pm \frac{1}{2})\Delta x, (j \pm \frac{1}{2})\Delta y, (k \pm \frac{1}{2})\Delta t)$ are obtained by linear interpolation. Developing and rearranging the terms we obtain :

$$v_{i,j,k}^{n+1} = v_{i,j,k}^n \left(1 - \frac{\Delta\tau}{h^2}\sum_l D_l^n\right) + \frac{\Delta\tau}{h^2}\sum_l D_l^n v_l^n\,, \tag{12}$$

where $l = \{E, W, N, S, F, R\}$ are the mnemonic subscripts for East, West, North, South, Front, Rear, and :

$$D_E = D_{i+\frac{1}{2},j,k}\,, \quad v_E = v_{i+1,j,k}\,, \tag{13}$$

$$D_W = D_{i-\frac{1}{2},j,k}\,, \quad v_W = v_{i-1,j,k}\,, \tag{14}$$

$$D_N = D_{i,j+\frac{1}{2},k}\,, \quad v_N = v_{i,j+1,k}\,, \tag{15}$$

$$D_S = D_{i,j-\frac{1}{2},k}\,, \quad v_S = v_{i,j-1,k}\,, \tag{16}$$

$$D_F = D_{i,j,k+\frac{1}{2}}\,, \quad v_F = v_{i,j,k+1}\,, \tag{17}$$

$$D_R = D_{i,j,k-\frac{1}{2}}\,, \quad v_R = v_{i,j,k-1}\,. \tag{18}$$

Thus, at each point $(i\Delta x, j\Delta y, k\Delta t)$ and iteration $n+1$ the intensity of the video signal depends only on its previous intensity and the intensities of the six closest spatial neighbors at iteration n. The contribution of each spatial neighbor v_l is determined by the interpolated diffusion coefficient D_l. Concerning the rest of the studied boundary value problem, the original video signal is used as initial condition in equation (2), and the boundary condition in equation (3) has been accomplished by setting the diffusion coefficient D to zero at the video boundaries.

A choice of $\Delta\tau \in [0, 1/6]$ ensures the positiveness of all coefficients in equation (12) since $h = 1$ and $D \in [0, 1]$. Under those conditions, our discretization satisfies the maximum principle introduced in Sec. 2. This can be proven easily by extending from two to three dimensions the demonstration performed by Perona and Malik in [10]. Thus, if we define the maximum and the minimum of the neighbors of $v_{i,j,k}$ at iteration n as $v_M = \max\{(v, v_l)_{i,j,k}^n\}$ and $v_m = \min\{(v, v_l)_{i,j,k}^n\}$ for $l = \{E, W, N, S, F, R\}$, we can prove from equation (12) that :

$$(v_m)_{i,j,k}^n \leq v_{i,j,k}^{n+1} \leq (v_M)_{i,j,k}^n\,. \tag{19}$$

As a result, at each iteration the maximum and the minimum of v become closer and no new maxima or minima are created.

5. EXPERIMENTS

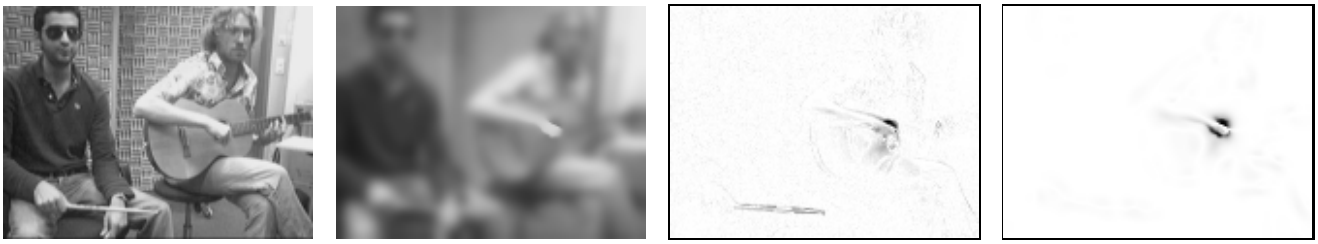
In all experiments the video signal $v(\mathbf{x}) \in [0, 255]$ and the audio activity $a(\mathbf{x}) \in [0, 1]$. The parameter $\Delta\tau = 0.15$ to satisfy the maximum principle in equation (19). The regularization parameter in (5) has been fixed to $\sigma = 0.3$. This value is enough to avoid noise and ensure video spatio-temporal coherence while providing a good spatial resolution. Videos showing the original and resulting video signals are available online at `http://lts2www.epfl.ch/~llagoste/AVdiff.htm`.

In **Sequence1**, a person is playing a guitar while another one is shaking a pair of drumsticks without making any sound (see Fig. 2(a)). The video is sampled at 30 fps with a resolution of 240×320 pixels, and the audio at 44 kHz. For its analysis, the video has been resized to 120×160 pixels, while the audio has been sub-sampled to 8 kHz. The sequence is $\approx$16s long. The parameter K has been fixed to the 95th percentile of s_σ ($K = 1.8$). Fig. 2(c) shows the original motion (computed as $|\delta_t^* v^0|$ in equation (11)) in a frame where the distracting motion caused by the drumsticks is important. After applying $n = 30$ iterations of our method, the frame is blurred and the only edges that remain sharp are situated around the guitarist's hand (see Fig. 2(b)). In the resulting motion map (Fig. 2(d)), all the distracting motion disappears and the guitarist hand is indicated as sound source. The robustness of our approach to severe Gaussian noise is also tested on this sequence. The quality of the video signals in the noisy case with respect to the clean case are expressed in terms of PSNR. The noise effect can be observed in Fig. 2(e) and specially in Fig. 2(g), where the real video motion is very difficult to distinguish between the noisy motion. Even in such a challenging situation our method is able to converge towards a very similar result in terms of intensity and motion (Fig. 2(f) and 2(h)), leading to global PSNRs of 44 dB and 45 dB respectively.

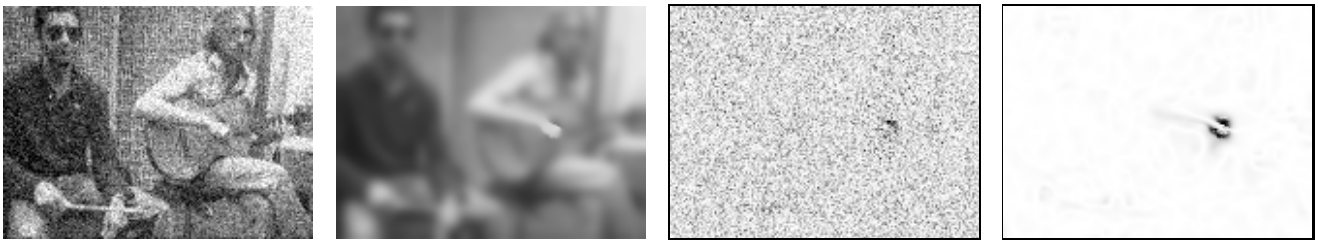
Sequence2 is taken from the state-of-the-art source localization work presented by Kidron et al. in [5]. It is composed of two moving objects: one of them is associated to the audio signal (a hand playing a guitar and then a synthetizer) and the other one represents a strong visual distraction (a rocking horse). The video is sampled at 25 fps 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, while the audio has been sub-sampled to 8 kHz. The sequence is $\approx$10s long. As expected, after $n = 34$ iterations of our method ($K = 1.5$) the hand that is playing the synthesizer is the only element well-defined in Fig. 3(b). Even if the rocking horse is moving continuously, its silhouette is blurred and its details are very difficult to appreciate. In Fig. 3(d) the relevant motion is kept while the rocking horse motion is attenuated. In fact, when the analyzed sequence contains a very consistent distracting motion for a long period, our algorithm is not able to remove it completely. In this case, our method keeps a slight focus on the rocking horse in temporal slots where its motion coincides with a sound because it *could* be the sound source.

6. DISCUSSION

We have introduced a novel method which is able to automatically smooth parts of a video signal that are not used in audio-visual analysis. Resultant sequences are thus simplified using an audio-based non-linear video diffusion. The main contribution of this work is the introduction of a diffusion coefficient which depends on an estimate

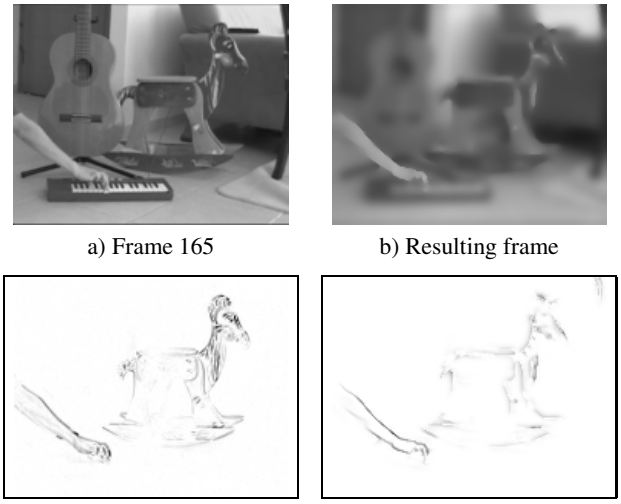


a) Clean frame 133 b) Resulting frame c) Original motion d) Resulting motion



e) Noisy frame PSNR = 20 dB f) Resulting frame PSNR = 44 dB g) Original motion PSNR = 18 dB h) Resulting motion PSNR = 45 dB

Fig. 2. Results obtained when applying our algorithm to **Sequence1** before [top] and after [bottom] adding Gaussian noise to the video signal.



a) Frame 165 b) Resulting frame

c) Original motion d) Resulting motion

Fig. 3. Results obtained when applying our method to **Sequence2**.

of the synchrony between video motion and audio energy at each point of the video signal. As a result, video regions presenting low coherence with the soundtrack are affected by homogeneous diffusion. The discretization of the proposed continuous diffusion model has been studied and its stability demonstrated.

Our method has been tested on sequences presenting video degradation and complex distracting video motion. Results show that sound sources are naturally highlighted after a few iterations, concentrating most of the resulting video motion. Our method is only based in the assumption of synchrony between audio and video channels, and thus distracting video motion which is consistent for a long period with the audio signal can not be completely eliminated. A post-processing step penalizing regions that move in silent periods could allow the application of our procedure to sound source localization. However, in this paper we want to keep our algorithm as general as possible so that other methods can profit from it.

7. REFERENCES

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

[2] Z. Barzelay and Y. Y. Schechner. Harmony in motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[3] F. Catte, P. Lions, J. Morel, and T. Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM Journal on Numerical Analysis*, 29(1):182–193, 1992.

[4] R. Goecke, G. Potamianos, and C. Neti. Noisy audio feature enhancement using audio-visual speech data. In *Proc. IEEE ICASSP*, 2002.

[5] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In *Computer Vision and Pattern Recognition (CVPR)*, 2005.

[6] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval. Blind Audiovisual Separation based on Redundant Representations. In *Proc. IEEE ICASSP*, 2008.

[7] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval. Learning Multi-Modal Dictionaries. *IEEE Trans. Image Process.*, 16(9):2272–2283, 2007.

[8] L. Nirenberg. A strong maximum principle for parabolic equations. *Comm. Pure Appl. Math.*, 6:167–177, 1953.

[9] H. J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: An empirical study. *CIVR*, pages 488–499, 2003.

[10] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(7):629–639, 1990.

[11] M. Siracusa and J. Fisher. Dynamic dependency tests: Analysis and applications to multi-modal data association. In *Proc. AIStats*, 2007.

[12] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz. Developing an audio-visual speech source separation algorithm. *Speech Commununication*, 44(1-4):113–125, 2004.

[13] J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner, 1998.