Geometric Video Approximation Using Weighted Matching Pursuit

Òscar Divorra Escoda[‡], Gianluca Monaci[*], Rosa M. Figueras i Ventura[**],
Pierre Vandergheynst and Michel Bierlaire[†]
Signal Processing Laboratory 2
Ecole Polytechnique Fédérale de Lausanne (EPFL)
LTS2-SPLabs-STI-EPFL, station 11, 1015 Lausanne, Switzerland
e-mail: oscar.divorra@ieee.org, gianluca.monaci@philips.com, rosa.figueras@ieee.org
{pierre.vandergheynst,michel.bierlaire}@epfl.ch
phone: +41 21 693 56 45, fax: +41 21 693 76 00

***Abstract*— In recent years, works on geometric multidimensional signal representations have established a close relation with signal expansions on redundant dictionaries. For this purpose, Matching Pursuits (MP) have shown to be an interesting tool. Recently, most important limitations of MP have been underlined, and alternative algorithms like Weighted-MP have been proposed. This work explores the use of Weighted-MP as a new framework for motion-adaptive geometric video approximations. We study a novel algorithm to decompose video sequences in terms of few, salient video components that jointly represent the geometric and motion content of a scene. Experimental coding results on highly geometric content reflect how the proposed paradigm exploits spatio-temporal video geometry. 2D Weighted-MP improves the representation compared to those based on 2D MP. Furthermore, the extracted video components represent relevant visual structures with high saliency. In an example application, such components are effectively used as video descriptors for the joint audio-video analysis of multimedia sequences.**

***Index Terms*— Video Representation, Spatio-temporal Decompositions, Geometry, Sparse Approximations, Redundant Dictionaries, Wavelets, *A Priori* Knowledge, Weighted Matching Pursuit.**

I. Introduction

Visual data representation has been a very active field of research in recent years. Many works have underlined the importance of compact representations of image and video signals for a wide range of applications such as compression, denoising, signal analysis or data mining, where a well defined and structured decomposition of data is required. Good modeling of visual data implies to capture its main structural properties, exploiting as much as possible underlying correlations. Moreover, models need to be compact (or sparse) such that dimensionality is reduced as much as possible.

In this direction, geometry adaptive representations have been found to be of capital importance in order to properly exploit the underlying structure of images. For this purpose, image models based on the additive superposition of geometric primitives (or atoms) have been proposed and studied, typically based on over-complete bases (e.g. [2], [3], [4]), or general highly coherent[1] dictionaries [5].

Concerning video representation, models based on the additive superposition of 3D functions have been typically related to separable 3D wavelet transforms [6] or to motion compensated variants of these [7], [8]. However, such models do not specifically take into account the geometric structure of video signals.

Most recently, works such as [9],[10],[11],[12] investigate the use of more geometry related video models. These works try to develop an intuitive video signal representation framework that takes into account the underlying geometric nature of spatial video components. Considering the fact that such components may have a transformation through time, video sequences can thus be expressed as a linear superposition of 3D functions representing spatial

Web page: http://lts2www.epfl.ch
This work was supported by the Swiss Federal Office for Education and Technology under grant number 6044.1 KTS. The work of Gianluca Monaci has been supported by the Swiss NFS through the IM.2 National Center of Competence for Research.

[‡] Òscar Divorra Escoda performed this work in the framework of his PhD Thesis [1] at the Signal Processing Institute (ITS), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Now, he is with Telefonica Research, Barcelona, Spain

[*] Gianluca Monaci contributed to this work while at the Signal Processing Institute (ITS), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Now, he is with Philips Research, Eindhoven, The Netherlands

[**] Rosa M. Figueras i Ventura contributed to this work while at the Signal Processing Institute (ITS), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Now, she is with the Group of Interactive Coding of Images at Universitat Autonoma de Barcelona, Catalonia, Spain.

[†] Dr. Michel Bierlaire was with the Mathematics Institute (IMA), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Now, he is with the Transport and Mobility Laboratory also at EPFL.

[1] The coherence is a measure often used to characterize redundant dictionaries which is defined as $max_{i,j|i \neq j} |\langle g_i, g_j \rangle|$, where g_i and g_j for any i, j are elements of the dictionary. Through the paper, highly coherent dictionaries stands for redundant dictionaries with a high cross-correlation among elements.

geometry structures and their evolution through time as:

$$\hat{f} = \sum_{k=0}^{K-1} c_k g^{3D}_{\gamma_k}, \quad (1)$$

where $\hat{f}$ is an approximation of a video signal, $g^{3D}_{\gamma_k}$ are 3D video atoms (s.t. $\forall \gamma_k \; g^{3D}_{\gamma_k} \in \mathcal{D}$, where $\mathcal{D}$ is a dictionary of functions) and c_k are scalars weighting the different spatio-temporal components of the model.

Signal models such as (1) typically require to be sparse in order to supply efficient and compact signal descriptions. To achieve sparseness, the large variety of components present in natural video signals have to be represented using adapted sets of basis functions that can efficiently describe them. Akin to images, but with the added complication of temporal deformation, large, redundant, and possibly coherent, dictionaries are required for sparse video approximations. Nevertheless, one cannot consider directly using such a large dictionary for video decomposition due to complexity issues. Instead, a possibility is to split into two steps the extraction of spatial and temporal components in order to lessen computational requirements. For this purpose, an approach is to first find spatial geometric components in a particular frame, and then track these through time. Hence, each $g^{3D}_{\gamma_k}$ can be represented by means of a parametric description of spatial geometric properties, plus the changes of these through time.

This paper studies the use of greedy algorithms for such a two step decomposition approach. First we consider the use of Matching Pursuit (MP) for the retrieval of spatial 2D components (as proposed in [5]). Then, a similar decomposition approach is used in order to recover the changes of 2D functions through time, i.e. trying to match each 2D atom from a frame at time t with its corresponding one at time $t+1$. However, MP is unsuitable for retrieving such spatio-temporal correspondences when dictionaries in use are too coherent [13], because it falls in local minima that are far from the optimal solution. Based on the findings of [14], [15], Weighted Matching Pursuit[2] (Weighted-MP) is used instead in order to overcome MP limitations. Weighted-MP considers the use of *a priori* knowledge within the signal decomposition process. This makes possible to take into account more complex and rich signal models that consider additional features such as joint motion/signal structure and the interaction among 3D atoms.

In this paper, we formalize a new geometric video representation framework. As assessed by the results, Weighted-MP is able to recover more compact decompositions (such as (1)) than MP, reducing the negative effects of dictionary coherence. Our approach shows to provide a signal representation with relevant structural and semantic content. In the results, examples are also shown on the potential possibilities of the studied representation for geometry based video compression, as well as a source of video features in an audio-visual source localization application.

This paper is organized as follows: Image modeling based on the superposition of geometric 2D components is recalled in Sec. II, along with a proposal for the extension of such a modeling framework to video signals. Then, Sec. III reviews MP and Weighted-MP algorithms, together with a general comparison between them. A video decomposition approach based on Weighted-MP and the use of *priors* on video and motion structures is presented in Sec. IV. Sec. V shows results on the use of Weighted-MP and MP for geometry-adaptive spatio-temporal video decompositions. Finally, conclusions are drawn in Sec. VI.

II. A Geometry-Adaptive Video Model

The proposed video model, together with Weighted-MP, builds upon the works [16], [17], [5], [18], [19], where geometry-adaptive image models were investigated using geometric redundant dictionaries and MP. In the following, we first recall the image model described in these works; then, we discuss how this model can be extended to approximate video signals.

A. A 2D Geometry-Adaptive Image Model

Images are often represented or approximated as a finite sum of 2D basis functions:

$$\hat{f} = \sum_{k=0}^{K-1} c_k g_{\gamma_k}, \quad (2)$$

where $\hat{f}$ is the image approximation, c_k are the coefficients and g_{γ_k} the selected functions ($\forall \gamma_k \; g_{\gamma_k} \in \mathcal{D}$), being $\mathcal{D}$ a dictionary of waveforms.

In order to have efficient image representations, basis functions have to adapt to contours, smooth regions and textures. Contours are often the most relevant and meaningful feature in natural images. Hence, they deserve the use of appropriate, adapted dictionaries of functions to efficiently describe them. Contours are assumed to be 1D continuous smooth functions [20], [21], [22], [4], [2], [17] with high geometric meaning.

Geometric dictionaries used in these are generated by applying a set of transformations to a mother function g [5], [18], [19]. The dictionary is spanned by a family of unitary operators $U(\gamma)$:

$$\mathcal{D} = \{U(\gamma)g, \gamma \in A\}, \quad (3)$$

for a given set of geometric transformations A. g defines the structural and morphological properties of the dictionary functions, while A captures most of its geometrical properties. In the remaining of the paper $g_\gamma = U(\gamma)g$.

Here we adopt the mother function g [5]:

$$g(u,v) = C(4u^2 - 2)\exp\left(-(u^2+v^2)\right), \quad (4)$$

where C is a normalizing constant and (u,v) are, in this case, coordinates on a plane. In (4), the v axis is formed by a Gaussian function that has the capacity to represent

[2] Weighted Matching Pursuit was first proposed as an heuristic solution in [14].

smooth structures. The perpendicular direction, is formed by a *Mexican Hat* (or *Marr*) wavelet [23], [24]. This intends to represent big variations within the signal, such as lines and edges.

Operator $U(\gamma)$ in (3) adapts (4) to the necessary variety of scalings, rotations and translations:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \frac{1}{s_x} & 0 \\ 0 & \frac{1}{s_y} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x-d_x \\ y-d_y \end{bmatrix}, \quad (5)$$

i.e. each γ index identifies a set of parameters $\{\mathbf{d}, \theta, s_x, s_y\}$. s_x and s_y define the anisotropic scaling of the function, θ sets the rotation of the function and finally, $\mathbf{d} = (d_x, d_y)$ are the translation of the function. (x, y) denote the image coordinates. In order to avoid ill-formed functions inappropriate to edges, only atoms with $s_y \geq s_x$ are taken into account (edges require long smooth functions along the contour with thin oscillating structures perpendicular to the contour).

A very low resolution version of the image is also used as part of the components in (2) in order to represent non-zero mean smooth image components.

B. A 3D Geometry-Adaptive Video Model: Tracking 2D Image Features Through Time

In the framework of video, 2D geometric features often follow temporal geometric transformations. A way to represent this is by modeling video as a superposition of 3D primitives that jointly capture spatial geometry and temporal evolution. However, most video representation paradigms separate motion information from image structures. This may produce less compact signal models, having a negative impact on the performance and/or efficiency of video applications.

Dictionaries of 3D spatio-temporal geometric functions may have a size that is difficult to deal with. Even for simple MP, this may turn atoms search into an intractable task. Moreover, temporal geometry evolution is often so complex that a dictionary, able to represent them, is likely to be extremely coherent. It is, thus, necessary to adopt strategies that take into account some *prior* about the signal in order to: i) simplify the search problem and ii) reduce the impact of dictionary coherence on MP.

Let I_t $\forall t$ be a set of consecutive images in a sequence. The changes suffered from frame I_t to I_{t+1} can be modeled as the application of an operator F_t on the image I_t such that

$$\begin{aligned} &I_{t+1} = F_t\left(I_t\right), \\ &I_{t+2} = F_{t+1}\left(I_{t+1}\right) = F_{t+1}\left(F_t\left(I_t\right)\right), \\ &I_{t+3} = \ldots \end{aligned} \quad (6)$$

where the subindex t indicates time.

From (2) and (6), $\hat{I}_{t+1}$ is modeled as a transformation of the geometric representation of $\hat{I}_t$ (where $\hat{I}_t$ stands for an approximation of I_t):

$$\hat{I}_{t+1} = F_t\left(\sum_{\gamma\in\Gamma} c_\gamma^t \cdot g_\gamma^t\right). \quad (7)$$

A relation needs to be established between F_t and the transformation of each one of the 2D components involved in the frame approximation. This is why we make the hypothesis that F_t is composed by the set of F_t^γ that independently transform each one of the frame expansion terms, i.e.

$$\hat{I}_{t+1} = \sum_{\gamma\in\Gamma} F_t^\gamma\left(c_\gamma^t \cdot g_\gamma^t\right). \quad (8)$$

No simple global model can be established for the joint transformation of all geometric primitives. Hence, local transformation models, applied to the basic elements used for frame approximation g_γ^t, are considered.

In the following, F_t may sometimes be referred to as a *deformation*. The action of each F_t^γ in (6) corresponds to a geometric deformation on g_γ and to a change of coefficient c_γ^t. Intuitively, this mechanism intends to implement local change of scale (s_x, s_y), position (d_x, d_y) and/or orientation (θ) for each primitive (see Fig. 1(a) and 1(b)). The sequence of deformations $F_t^\gamma : t \in [T_1, T_2]$ and the 2D atom g_γ^t form a 3D primitive that represents how scene geometry flows through time.

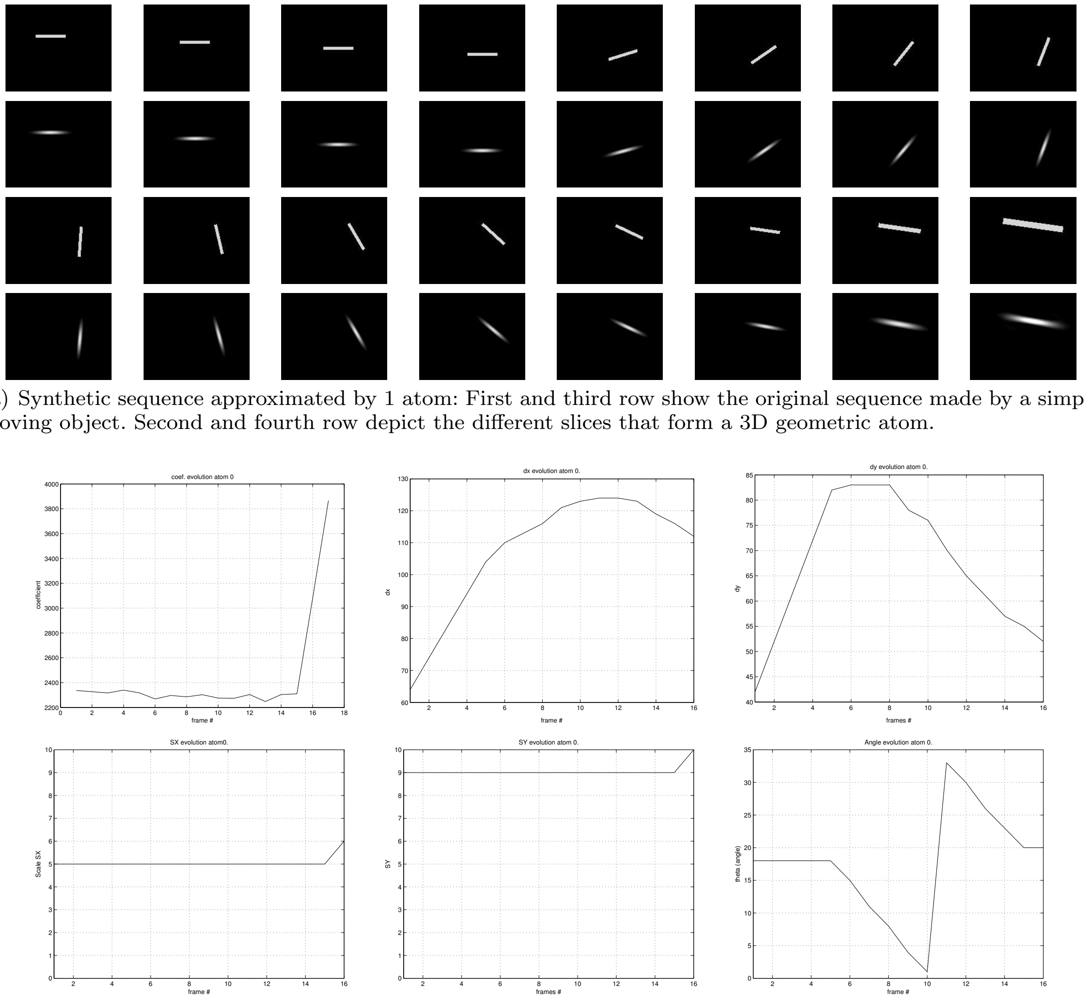
In this work, F_t^γ operators are approximated as the set of differential parameter values $\Delta\mathbf{d} = (\Delta d_x, \Delta d_y)$, $\Delta\mathbf{s} = (\Delta s_x, \Delta s_y)$, $\Delta\theta$ mapping atom g_γ from time t to $t+1$:

$$\forall\gamma,\ \forall t \quad g_\gamma^t \xmapsto{F_t^\gamma} g_{\gamma'}^{t+1} \text{ s.t. } g_\gamma^t, g_{\gamma'}^{t+1} \in \mathcal{D}. \quad (9)$$

In order to be consistent with 2D pictures representation, we impose that F_t^γ is such that g_γ^t and $g_{\gamma'}^{t+1}$ belong to the same dictionary $\mathcal{D}$. This allows the use of fast atom search algorithms (such as the one used for 2D image approximation [25]). Since the set of F_t^γ are used as a parametric description for coding purposes, a quantization on the evolution of geometric parameters γ is required. Such a quantization needs to be considered within the decomposition loop of the greedy algorithm [9]; hence, deformed atoms must also belong to $\mathcal{D}$.

The set of all possible transformations F_t^γ is an approximation of the affine model of local transformations defined for sequences. This approximation intends to supply a trade off between adaptation flexibility and dictionary complexity. The model considered in this work does not include shearing and is limited by the granularity of $\mathcal{D}$ parameters. According to (6), geometric video primitives are recovered on a frame by frame basis, tracking its deformation through time to recover the video structures.

A graphic example can be seen in Fig. 1(a), where the approximation of a simple synthetic object by means of a single atom is performed. First and third picture rows show the original sequence and the second and fourth rows provide the reconstruction of the approximation. Fig. 1(b) shows the parametric representation of the sequence. We see the temporal evolution of the coefficient c_γ^t, and the other 2D geometric parameters.



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

(b) Parameter evolution of the approximated object; from left to right and from up down, we find: coefficient, x position, y position, x (short axis) scale, y (long axis) scale, angle.

Fig. 1. Approximation of a synthetic scene by means of a 3D atom.

III. Weighted Matching Pursuit

A. A General Greedy Algorithm: Matching Pursuit

Sparse approximation of signals using highly redundant dictionaries typically requires the use of sub-optimal strategies due to computational feasibility issues [26], [27], [28]. One of the most popular algorithms, due to its simplicity and usability with highly redundant dictionaries, is the so-called Matching Pursuit (MP) [27].

Consider an over-complete dictionary $\mathcal{D}$ where atoms belong to $\mathbb{R}^N$. General MP [29], [30], [27] iteratively builds m-term approximants by selecting at each step the most appropriate term from $\mathcal{D}$ according to a certain rule. Each one of these iterations can be seen as a two step procedure:

1) A selection step where an atom $g_{i_k} \in \mathcal{D}$ is chosen (where $k \geq 0$ indicates the iteration number).
2) A projection step where an approximant $f_m \in span(g_{i_k} : k \in \{0, ..., m-1\})$ and a residual $r_m = f - f_m$ are generated.

The selection step, at iteration k, can be formulated as the maximization of a similarity measure $C(r_k, g_i)$ between the signal to approximate (the residual at the kth iteration: $r_k = f - f_k$) and the dictionary atoms:

$$g_{i_k} = \arg\max_{g_i \in \mathcal{D}} C(r_k, g_i). \tag{10}$$

MP uses the modulus of the scalar product as similarity measure in order to minimize the projection error, i.e. $C(r_k, g_i) = |\langle r_k, g_i \rangle|$.

Consider $r_0 = f$; then, at the first iteration, MP will represent the signal as:

$$f = r_0 = \langle f, g_{\gamma_0} \rangle g_{\gamma_0} + r_1 \,, \tag{11}$$

where r_1 is the residual after approximating r_0 in the direction of g_{γ_0}.

Hence, from (11):

$$f = \sum_{k=0}^{K-1} \langle r_k, g_{\gamma_k} \rangle g_{\gamma_k} + r_K. \tag{12}$$

Recent studies like [13], [29] suggest that the use of incoherent dictionaries is very important for the good behavior of MP. However, as underlined in [15], experience teaches us that highly coherent dictionaries are more powerful for natural signal approximations [5], [2], [21]. Alternatives to

MP are thus required for good sparse decompositions when highly overcomplete (and coherent) dictionaries are in use. Yet, usable alternatives must be simple enough.

B. Using A Priori *Information within Matching Pursuit: Weighted Matching Pursuit*

The use of the scalar product as similarity measure in MP bears some similarity with searching for the "Most Likely" atom g_{i_k} given the residual r_k: i.e. the atom g_{i_k} that maximizes the probability $p\left(g_i|r_k\right)$ is selected. Thus, $|\langle r_k, g_i\rangle|$ may be intuitively seen as a measure of the conditional probability $p\left(r_k|g_i\right)$, and when all g_i are equally probable, maximizing $|\langle r_k, g_i\rangle|$ is equivalent to maximizing $p\left(g_i|r_k\right)$. Based on this, [15], [14], [1] consider the case where, depending on the signal, dictionary waveforms do not have the same *a priori* probability to appear in the optimal set of m atoms (Γ_m). Indeed, the authors propose an algorithmic approach where some *prior* knowledge $p\left(g_i\right)$ on the likelihood of each g_i is available. By the Bayes' Rule, the probability to maximize becomes

$$p\left(g_i|r_k\right)=\frac{p\left(r_k|g_i\right)p\left(g_i\right)}{p\left(r_k\right)}, \tag{13}$$

where the denominator is usually assumed to be constant for any signal r_k. Based on this, one can modify the MP selection rule by multiplying the modulus of the scalar product by a weighting factor $w_i \in (0,1]$ (which depends on the atom index i) in order to represent the insertion of a measure of *prior* information. Hence, now $C(r_k, g_i)$ in (10) is such that:

$$C\left(r_k, g_i\right)=|\langle r_k, g_i\rangle|\cdot w_i. \tag{14}$$

See [1] for a formal relation between w_i and $p\left(g_i\right)$.

The general family of weighted greedy algorithms using (14) are called Weighted-MP [14], [15]. Weighted-MP does not modify the projection step of the algorithm, allowing to freely select MP or OMP (Orthogonal MP). *A priori* knowledge ($w_i\ \forall i \in \Omega$) may change, if necessary, depending on the iteration k and the signal to approximate as shown in Alg. 1.

Algorithm 1 : Weighted-MP flow

```
Compute Initial A priori information (generate $w_i\ \forall i$).
Initialize $r_0 = f$, where $f$ represents the original signal
  and $r$ is the residual signal to approximate. Initialize
  $\hat{f} = 0$, where $\hat{f}$ is the generated approximation.
for $k = 0$ to K-1 do
    Find $g_{i_k}$ s.t. $g_{i_k} = \arg\max_{g_i \in \mathcal{D}} C(r_k, g_i)$, where
      $C\left(r_k, g_i\right) = |\langle r_k, g_i\rangle| \cdot w_i$
    $r_{k+1} = r_k - \langle r_k, g_i\rangle \cdot g_{i_k}$
    Update the probability maps considering $g_{i_k}$ in or-
      der to generate the new $w_i\ \forall i$ for the next iteration.
    Generate approximation $\hat{f}_{k+1} = \hat{f}_k + \langle r_k, g_i\rangle \cdot g_{i_k}$
end for
```

As demonstrated in [15], when coherent dictionaries are used, Weighted-MP is more likely to recover good signal approximations than MP [13], [29]. This helps to better respect the structural nature of signal components. At the same time, it also proves that a more sophisticate atom selection criterion help to speed-up the convergence of greedy approximations, which reduces the negative effect of dictionary coherence on the greedy algorithm.

In this work, the use of Weighted-MP in the framework of geometry-adaptive video approximations is investigated. A geometric video model is proposed in the next sections for both: video representation, and as an *a priori* model to drive Weighted-MP. The use of Weighted-MP and the *prior* is compared with simple MP by considering diverse application scenarios.

IV. A Geometry-Adapted Video Decomposition Based on Weighted-MP

In this section, we propose an *a priori* model in the selection step of Weighted-MP for frames that have temporal precedent frames. The first frame of a sequence or group of pictures is decomposed using a full MP as described in Sec. II, and the rest of the frames, using weighted MP.

A. Weighted-MP Selection Criteria for Video Decomposition

Video frames decomposition can be formulated from a Bayesian point of view if some *a priori* information about F_γ^t, the relation between atoms depending on γ and/or their temporal evolution, is available. For this purpose, we make the assumption that neighboring atoms, in space and time, present similar deformations. This can be done by the use of proper regularization terms in (13); i.e. by using a Bayesian functional in (10) that maximizes the Maximum a Posteriori (MAP) probability. In the following, Weighted-MP selection criterion is formulated considering a Markovian framework in order to define probabilistic relations among atom motions.

For every Weighted-MP iteration, in the decomposition of a video frame the following expression is optimized:

$$\begin{aligned} F_t^\gamma = \Delta\gamma_n &= \arg\max_{\Delta\gamma_n}\left\{p\left(\Delta\gamma_n, \Delta c_n \mid r_n^{t+1}, g_{\gamma_n}^t\right)\right\} \\ &= \arg\max_{\Delta\gamma_n}\left\{p\left(r_n^{t+1}, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right)\cdot p\left(\Delta\gamma_n, \Delta c_n\right)\right\}, \end{aligned} \tag{15}$$

where $\Delta\gamma_n$ represents the parameter differences between ${\gamma'}_n^{t+1} \in \Gamma$ and $\gamma_n^t \in \Gamma$, and r_n^{t+1} is the nth iteration frame residual at time $t+1$.

Equation (15) makes use of the Bayes' rule to establish the atom selection criterion, by maximizing the matching probability of a given $g_{\gamma_n}^t$ with the residual r_n^{t+1}, conditioned to the probability of the transformation $\Delta\gamma_n$ and the temporal change on the projection coefficient Δc_n. The matching probability $p\left(r_n^{t+1}, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right)$ can be defined as a function of an estimated residual error energy $\left\|\hat{r}_{n+1}^{t+1}\right\|^2$ for the retrieval of function g_{γ_n} at iteration n. Atoms are assumed to deform through time under consistent motion transformation. Thus, the coefficient should not change (except for scale changes) in the estimation of the most probable motion. Hence,

$$\hat{r}_{n+1}^{t+1} = {r_n}^{t+1} - \overline{\left\langle r_n^t, g_{\gamma_n}^t \right\rangle} g_{\gamma'_n}^{t+1}, \tag{16}$$

where $\overline{\left\langle r_n^t, g_{\gamma_n}^t \right\rangle}$ stands for the scalar product normalized according to the re-scaling of $g_{\gamma'_n}^{t+1}$ with respect to $g_{\gamma_n}^t$, i.e $\overline{\left\langle r_n^t, g_{\gamma_n}^t \right\rangle} = \left\langle r_n^t, g_{\gamma_n}^t \right\rangle \sqrt{\Delta s_x \Delta s_y}$, where $\sqrt{\Delta s_x \Delta s_y}$ renormalizes the new atom to preserve unitary norm. The re-scaling is necessary in order to compensate for temporal changes in scale of the normalized atoms, so that the amplitude of the data they represent is maintained. At time t, $r_{n+1}^t \perp g_{\gamma_n}^t$, in order to minimize the energy of the projection error. In the same way, after motion transformation, $g_{\gamma'_n}^{t+1}$ should be such that $\left\|\hat{r}_{n+1}^{t+1}\right\|^2$ is also minimized.

The probability measure assumes Gaussianity (by the central limit theorem [31]) and independence of error samples $r_{n+1}^t(x,y)$ for simplicity. Akin to other approaches such as block matching and Markov Random Fields (MRF) [32], [33], we consider:

$$\begin{gathered} p\left(r_n^{t+1}, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right) = \\ \frac{1}{Z} \prod_{x,y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left|\hat{r}_{n+1}^{t+1}(x,y)\right|^2}{2\sigma^2}\right) \end{gathered} \tag{17}$$

where Z is a normalizing constant and $\sigma^2 \approx E\left[\left|\hat{r}_{n+1}^{t+1}(x,y)\right|^2\right]$. Note that $\hat{r}_{n+1}^{t+1}$ is considered to have zero mean. In fact, prior to any operation, a low pass approximation is removed from every frame (as discussed in Sec. II). Introducing the evaluation of σ^2 in (17) we obtain the conditioned optimization criterion:

$$p\left(r_n^{t+1}, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right) \approx \frac{C_1}{\sqrt{\left\|\hat{r}_{n+1}^{t+1}\right\|^2}}, \tag{18}$$

where C_1 is a constant.

Probability $p\left(\Delta\gamma_n, \Delta c_n\right)$ imposes the model that drives the transformation F_t^γ of the g_γ^t and the associated coefficient. This is, thus, defined as the conditioned probability of $\Delta\gamma$ and Δc_n in the Markovian framework. At every iteration, Weighted-MP tries to select a new atom that maintains regularity with all those previously selected in the neighborhood (Fig. 2). Earlier atoms are trusted to generate the motion regularity estimates for future appearing atoms (within the same frame). This unbalanced criteria derives from the fact that first atoms of the Weighted-MP decomposition capture more energy, thus they tend to represent much more significant (i.e. reliable) signal features.

We can formulate $p\left(\Delta\gamma_n, \Delta c_n\right)$ as:

$$\begin{gathered} p\left(\Delta\gamma_n, \Delta c_n\right) = \\ p\left(\Delta c_n \mid \Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\right) \cdot p\left(\Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\right), \end{gathered} \tag{19}$$

where Δc_n (temporal variation of the nth atom scalar product with the residual) depends on the choice of new γ parameters. Considering $\Delta\mathbf{d}$, $\Delta\mathbf{s}$, $\Delta\theta$ independent, (19) turns into:

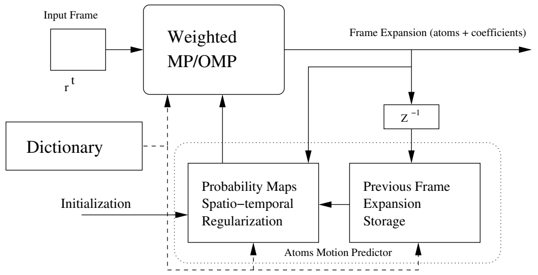

Input Frame
r t
Weighted MP/OMP
Frame Expansion (atoms + coefficients)
Z −1
Dictionary
Initialization
Probability Maps Spatio−temporal Regularization
Previous Frame Expansion Storage
Atoms Motion Predictor

Fig. 2. Expansion Block Scheme.

$$\begin{gathered} p\left(\Delta\gamma_n, \Delta c_n\right) = \\ p\left(\Delta c_n \mid \Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\right) \cdot p\left(\Delta\mathbf{d}_n\right) \cdot p\left(\Delta\mathbf{s}_n\right) \cdot p\left(\Delta\theta_n\right). \end{gathered} \tag{20}$$

Each of the probability functions has the form of a MRF, and thus, they may be modeled by a Gibbs distribution [34]:

$$p(x) = \frac{1}{Z_x} \exp\left(-\frac{E_x\left(x\right)}{T_x}\right), \tag{21}$$

where $E_x\left(x\right)$ is an energy function that characterizes the MRF and how neighboring variables are related, while T_x stands for its variance.

From (15), (18), (20) and (21), the functional to be optimized can be expressed as:

$$\begin{gathered} \Delta\gamma_n = \underset{\Delta\gamma_n}{\arg\min} \left\{ \frac{1}{2} \log\left(\left\|\hat{r}_{n+1}^{t+1}\right\|^2\right) + \lambda_{\Delta c_n} E_{\Delta c_n}\left(\Delta c_n\right) + \right. \\ \left. \lambda_{\Delta\mathbf{d}_n} E_{\Delta\mathbf{d}_n}\left(\Delta\mathbf{d}_n\right) + \lambda_{\Delta\mathbf{s}_n} E_{\Delta\mathbf{s}_n}\left(\Delta\mathbf{s}_n\right) + \lambda_{\Delta\theta_n} E_{\Delta\theta_n}\left(\Delta\theta_n\right) \right\} \end{gathered} \tag{22}$$

where $\Delta\gamma_n = \left\{\Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\right\}$ and each λ_x are a function of the statistics parameter T_x in (21).

B. Regularity Models

1) Coefficient Model: Temporal variations of coefficients, Δc_n, should be small in ideal tracking of an atom. In any case, coefficients may not change sign. Changes to coefficients should be driven mainly by the change of scale of the approximating function. A normalized quadratic distance between the coefficients at time t and $t+1$ is considered for $E_{\Delta c_n}\left(\Delta c_n\right)$:

$$E_{\Delta c_n}\left(\Delta c_n\right) = \left(\frac{c_n^{t+1} - c_n^t \cdot \sqrt{\Delta s_x \Delta s_y}}{c_n^t \cdot \sqrt{\Delta s_x \Delta s_y}}\right)^2, \tag{23}$$

where previous c_n^t are re-normalized with respect to the scale transformation. One can observe that (23) is normalized in order to be independent of n.

2) Geometric Models: Displacement, change of scale and rotation constraints are measured as the euclidean distance between the value under test and the most likely

(ML) transformation estimated from previous Weighted-MP iterations at every image location. Hence,

$$
\begin{aligned}
E_{\Delta \mathbf{d}_n} &= \left(d_{x,n} - \hat{d}_{x,n}\right)^2 + \left(d_{y,n} - \hat{d}_{y,n}\right)^2 \\
E_{\Delta \mathbf{s}_n} &= \left(s_{x,n} - \hat{s}_{x,n}\right)^2 + \left(s_{y,n} - \hat{s}_{y,n}\right)^2 \\
E_{\Delta \theta_n} &= \left(\theta_n - \hat{\theta}_n\right)^2 ,
\end{aligned}
\tag{24}
$$

where $\hat{d}$, $\hat{s}$ and $\hat{\theta}$ correspond to the ML estimates (see Sec. IV-C for details about their calculation).

C. Motion and Probability Fields Estimation

ML motion estimates are computed considering all atoms that interact within a given region. In the example presented in this work, atoms have a localized support in space (see (4)). Even though (4) has not a strictly finite support, amplitude decay is fast enough such that atoms located sufficiently far away can be considered not to interact among them. Furthermore, the decay of the Gaussian envelop of (4) can also be considered as a hint of the fact that the strength of constraints (23) and (24) should increase the closer an atom is from another one, i.e. two atoms have a more coherent motion the closer they are.

1) λ_x Modeling: From (4), the atom envelop is a bivariate Gaussian with the same scaling dimensions (s_x, s_y) as the atom itself:

$$
\begin{aligned}
p_\gamma(u,v) &= K \exp\left(-\left(u^2+v^2\right)\right) \quad \text{s.t.} \\
u &= \frac{\cos\theta\,(x-d_x)+\sin\theta\,(y-d_y)}{s_x} \\
v &= \frac{-\sin\theta\,(x-d_x)+\cos\theta\,(y-d_y)}{s_y},
\end{aligned}
\tag{25}
$$

where K is a constant. This model is assumed to represent the influence law of the transformation of a given atom in a neighborhood. Thus, for any x, λ_x depend on the spatial location and are proportional to $p_\gamma(u,v)$. That is, the variance of the probabilities described in Sec. IV-A depends on the spatial location and decreases as a function of scale and with the distance to the center of an atom. Every λ_x is a constant (tuned in order to fit the deformation model) multiplied by the bivariate model of (25):

$$\lambda_x(x,y) = C_{\lambda_x} \cdot p_\gamma(u,v).$$

As one can observe, λ_x depends on the area of influence of each atom g_γ. In a given area, where more than one atom overlap, only the highest $p_\gamma(u,v)$ at position (u,v) is considered.

2) Motion Parameter Estimates: Motion parameters $\hat{d}_x, \hat{d}_y, \hat{s}_x, \hat{s}_y, \hat{\theta}$ in (24) are estimated from the preceding $n-1$ atoms of current frame expansion. They are the ML estimates according to the energy probability associated to each atom.

In fact, considering that a given frame energy can be represented as the sum of the square of the coefficients in a MP expansion:

$$\left\|I_{t+1}\right\|^2 = \sum_{n=0}^{\infty} \left|c_n\right|^2 , \tag{26}$$

we approximate the probability associated with the n*th* atom as a fraction of $\|I\|^2$

$$p(\gamma_n) = \frac{|c_n|^2}{\|I\|^2}. \tag{27}$$

The conditioned probability that a given atom contributes to spatial location (x,y) can be modeled through (25). Thus,

$$p(x,y \mid \gamma_n) = \frac{K}{\sqrt{s_x \cdot s_y}} \exp\left(-\left(u(x,y)^2 + v(x,y)^2\right)\right). \tag{28}$$

The motion parameters induced by atom $g^t_{\gamma_n}$ at point (x,y) have probability:

$$p(\gamma_n \mid x,y) = \frac{p(x,y \mid \gamma_n)p(\gamma_n)}{\sum_m p(x,y \mid \gamma_m)p(\gamma_m)}. \tag{29}$$

We can see that the summation in the above equation will only integrate those atoms close to position (x,y) (due to their amplitude decay -(25)-). Giving as example the case of the ML displacement ($\hat{\mathbf{d}}_n$ in (24)) at a given (x,y), we formulate it as the weighted average of all the transformations induced by all the atoms at a given point:

$$\hat{\mathbf{d}}_n = E\left\{\mathbf{d}_n \mid x,y\right\} = \sum_{k=0}^{n-1} \hat{\mathbf{d}}_k(x,y) \cdot p(\gamma_k \mid x,y). \tag{30}$$

The same applies to the other parameters, $\hat{\mathbf{s}}$ and $\hat{\theta}$.

Only the precedent $(n-1)$ available atoms are considered for the statistical measurements and calculations of the ML motion. When no reliable estimate of $p\left(\Delta\gamma_n\right)$ is available (i.e. when none of the $(n-1)$ atoms spatially overlaps with the nth atom), an initial value is generated by matching the region (patch) where the atom is supported in frame t with frame $t+1$. The cross-correlation (matching) of the zero mean and normalized versions of the patch and the frame that we want to approximate is used, i.e. the correlation between the normalized patch and the normalized frame is measured for every possible geometric transformation of the atom.

D. Atom Refresh

This work considers a forward mapping scheme where all atoms from frame at time t try to get matched in the frame at time $t+1$. This is not always possible and sometimes the atom will not be able to find at $t+1$ the feature it was representing at time t. Thus we measure the reliability of the predicted atoms motion as follows. At every new frame, the normalized scalar product of the transformed atom is compared with the projection of the atom in the first frame:

$$\left|\frac{\left\|c_n^{t+1}\right\|^2}{s_{x,n}^{t+1} s_{y,n}^{t+1}}\right| \geq \frac{\|c_n^0\|^2}{s_{x,n}^0 s_{y,n}^0} \cdot \delta, \qquad \delta \in (0,1] \tag{31}$$

If a significant drop in the scalar product is detected, the atom is canceled (the trajectory is not valid anymore). At the end of the projection process, those atoms that have been canceled are reintroduced in the frame through a full MP search.

In the present investigation, when the atom refresh is used, it is set such that a very small portion of atoms can be renewed at every frame (e.g. no more than three percent).

V. Experimental Study

A series of experimental tests are presented to asses the effect of Weighted-MP with respect to MP in the extraction of spatio-temporal, geometry-adaptive, video structures. In the following, experiments are realized considering a 2D dictionary sampling such that (according to Sec. II-A), d_x and d_y have pixel resolution, rotation θ is sampled every $\pi/36$ and s_x and s_y are uniformly distributed on a logarithmic scale from one up to a fourth of the size of the image, with a resolution of a half of octave. In what concerns temporal search for the recovery of the 3D atoms deformation, a maximum search window of $+/-30$ pixels, $+/-2$ scale samples and $+/-4$ rotation samples is considered. Unless the contrary is said, natural sequences are decomposed using the following settings in our experiments: $\lambda_{\Delta c_n} = 2.5{\cdot}10^{-4}$, $\lambda_{\Delta \mathbf{d}_n} = 1{\cdot}10^{-3}$,$\lambda_{\Delta \mathbf{s}_n} = 1.25 \cdot 10^{-4}$, $\lambda_{\Delta\theta_n} = 6.5 \cdot 10^{-3}$ and $\delta = 0.2$. In the current implementation, a search based on the use of FFTs is used in order to compute scalar products [25]. In the following, first, a set of experiments evidences the motion regularization effect of *a priori* information within Weighted-MP. Then, the entropy of the generated video representations is measured and used for comparison. A simple coding scheme is used on video representations generated in this work for illustration. Finally, an example on the use of present video representation for multi-modal video analysis is shown.

A. Regularization Effect of Weighted-MP on Geometric Video Structures

Fig. 3 shows, by means of a motion flow, the local deformation of atoms from frame 2 to 3 in the sequence *foreman* (see the 1st row of Fig. 6 for an extract of this sequence). The left image shows geometry deformations captured by the MP based algorithm. The right image shows a much more stable set of deformations obtained by Weighted-MP. Unlike the representation generated by MP, video atoms captured by Weighted-MP have a local motion closer to the real one. The same phenomenon is observed for sequence *motorway* (Fig. 4). *A priori* driven selection criteria of Weighted-MP succeeds in extracting much more stable geometric video primitives than MP. Only those atoms really concerned by moving objects present a temporal deformation. The weaker selection criteria of MP gets more easily destabilized, leading to a much shakier (i.e. noisier) video representation.

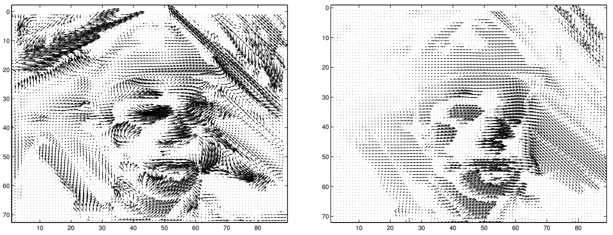


Fig. 3. Comparison of the computed deformations (atoms associated motion) for the 2nd frame of the *foreman* sequence: left not regularized, right regularized.

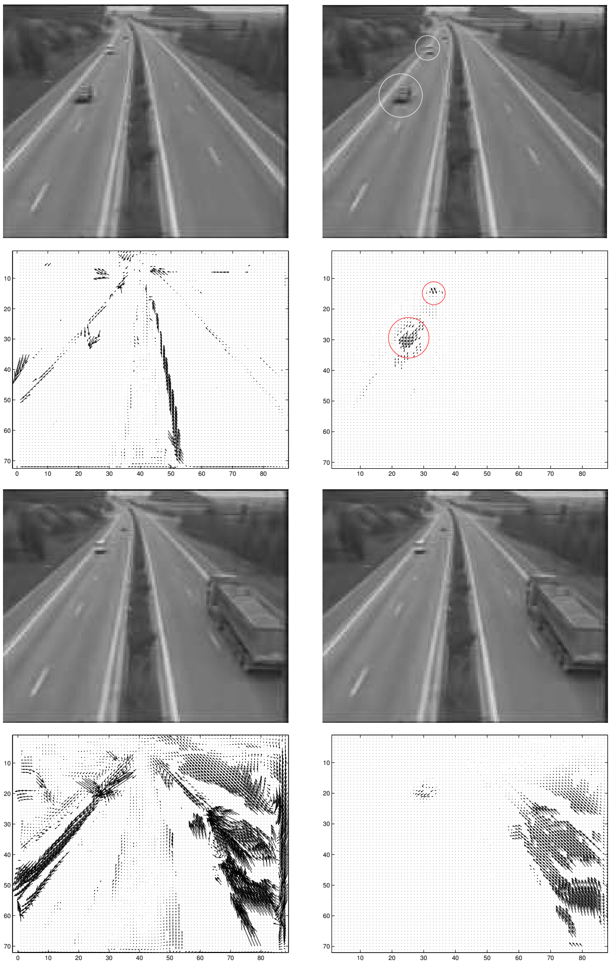


Fig. 4. Natural sequence *motorway*. Left column: non-regularized solution. Right column: regularized tracking. First and third rows: Respective reconstructions with 500 atoms. Second and fourth rows: Most reliable primitives motion.

In the following, two examples on the results generated by Weighted-MP are presented. For these results, $\lambda_{\Delta c_n} = 1.66 \cdot 10^{-2}$, $\lambda_{\Delta \mathbf{d}_n} = 0.0227 \cdot 10^{-2}$,$\lambda_{\Delta \mathbf{s}_n} = 4.0 \cdot 10^{-2}$, $\lambda_{\Delta\theta_n} = 0.129 \cdot 10^{-2}$. The first one can be seen in Fig. 5. This shows a sequence where a bar translates and rotates. In the second line, one can observe the motion flow of spatio-temporal atoms recovered by Weighted-MP.

Fig. 6 shows several geometric atoms with their evolution through time. These are represented in the second, third and forth row by showing the spatial "footprint" (in white) of a selected atom and its evolution through time

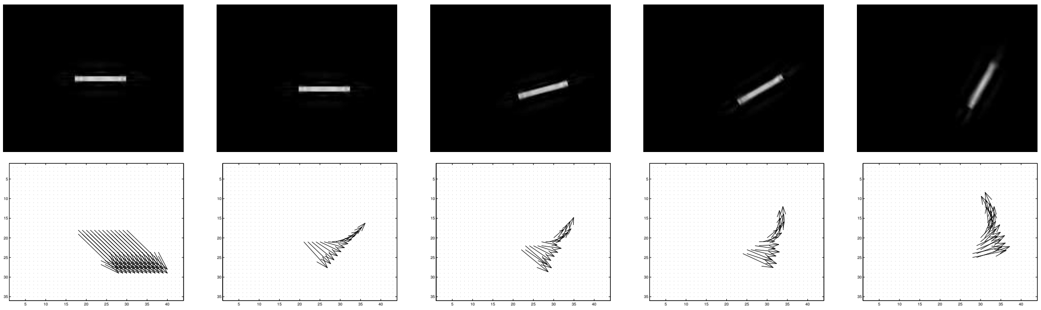


Fig. 5. Affine motion of a synthetic model (line). From top to bottom: approximation of the line, residual with respect to the original model and motion associated to the atoms. In the second row, we clearly see the effect of parameter quantization, in this case error is induced by the limited resolution in translations and rotations.

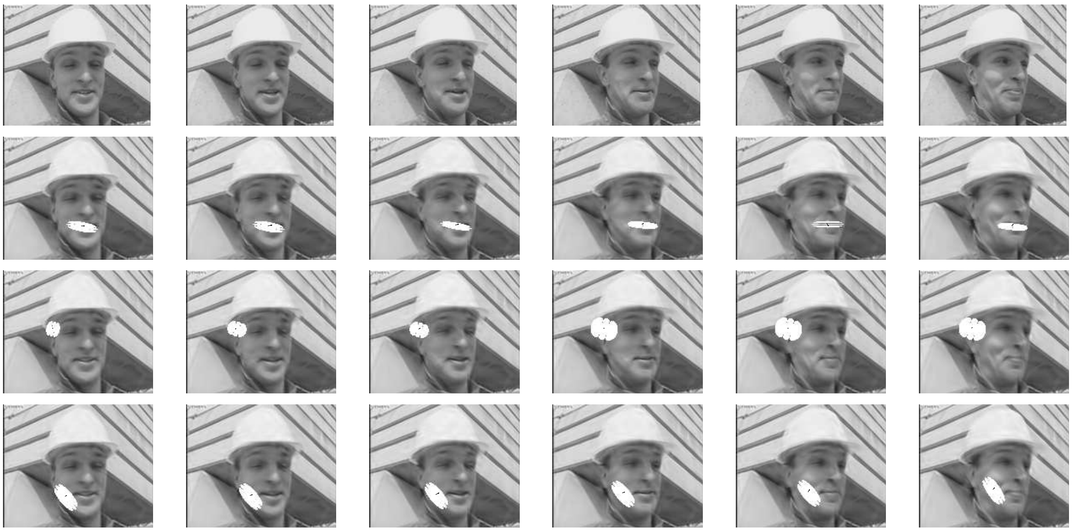


Fig. 6. Several consecutive frames of a natural sequence showing the reconstructed signal with 500 atoms. First row: the original frame; Second to Forth: motion of 3 different atoms from the sequence on the reconstructed signal. Their temporal evolution is indicated by the changes on the white footprint.

(from left to right). Combinations of different temporal geometric deformations such as translation, rotation and scaling can be observed in the three different examples displayed in Fig. 6.

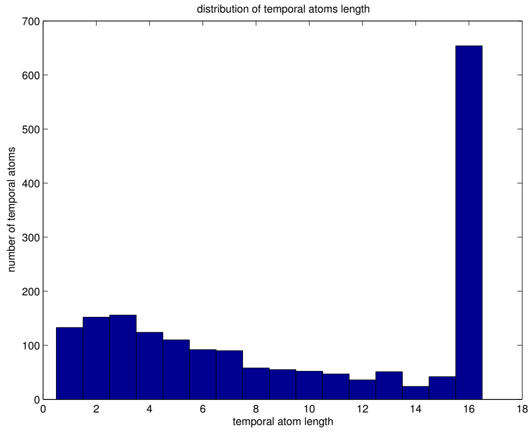


Fig. 7. Distribution of length for the temporal atoms. The length is determined by the atom refresh criteria of Sec. IV-D where atoms losing 80% of their amplitude are refreshed

Geometric video primitives do not necessarily last all along a GOP. Fig. 7 shows the histogram of temporal lengths for atoms prediction that are determined by the criteria described in Sec. IV-D. In this example the 48 first frames (3 GOPs) of the sequence *foreman* have been taken into account for the generation of the statistics. The total number of spatio-temporal atoms (sets of atoms that are predicted from frame to frame without being refreshed) within this 3 GOPs is 1876. There are about 35 per cent of atoms that succeed in being predicted from frame to frame during all the GOP. However, a relevant number needs to be refreshed quite often: common temporal lengths are from 1 to 8 frames. Sequence changes (occlusions, uncoverings and simple interaction among atoms) force their refresh. Atom refresh is a natural manner to introduce components to represent new information that appeared in the signal. This is the case in Fig. 8 (both pictures are decoded approximations reconstructed with 500 atoms each), where some regions on Foreman face get uncovered.

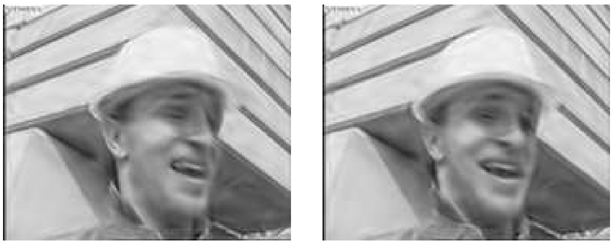


Fig. 8. Reconstructed frames 12,13 from the *foreman* sequence. In them, we observe the uncovering of the left region (right in the picture) of the man's face.

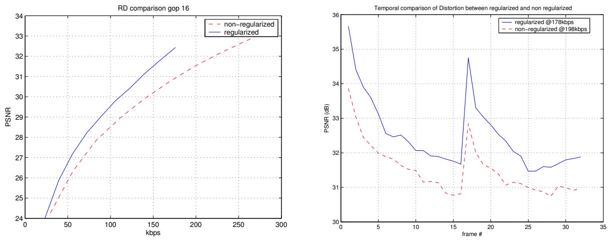


Fig. 9. Comparison of the regularized and non-regularized *foreman* sequences (16 frames GOP). Left: R-D, Right: Temporal comparison at two particular rates of the evolution of the PSNR with the frame number.

B. Entropic Effect of Weighted-MP on Geometric Video Structures

In here, a simple coding scheme is used to code a spatio-temporal video representation generated using the Weighted-MP video decomposition. This is used to evaluate the effect of using Weighted-MP instead of MP on a entropy basis. Tests are performed on Foreman sequence in QCIF format at 30Hz. For simplicity of the coding model, no atom refresh is considered in here; allowing to cumulate a more significant temporal drift with the frame number.

1) A Predictive Scheme for 3D Structures Coding: The coding scheme presented here tries to exploit the temporal redundancy of geometric video features with what has been presented in Sec. IV-D. The coding algorithm sees the video representation as a set of spatial atoms that are tracked through time. The tracking algorithm is the Weighted-MP video decomposition, which generates the set of data streams (coefficients and atom parameters changing through time).

Video representation data streams are, then, coded based on a DPCM [35] approach (in the present case, this is based on the simplest of the predictors, i.e. difference prediction, and a uniform dead-zone quantizer for the coefficients). Prediction residue is then coded by an adaptive m-ary arithmetic encoder [36], [37], [38]. Symbol statistics are independently estimated for each kind of parameter. A statistical context for every kind of parameter is reserved for "intra" coding (first DPCM sample), and another one is kept for "inter" coding (predicted samples). A module may estimate whether the trajectory of an atom is at its end. If this is the case, an additional signal could be transmitted to indicate that the atom, tracked until that point, is not tracked anymore and new intra data (new geometric atom) is introduced for tracking.

Predictive representation of spatio-temporal video components is performed on a limited length group of pictures (GOP) basis. A maximum prediction length (L) is fixed. Every L frames, atom trajectories are terminated and a new prediction GOP is started. In the present test, trajectories cover the full length of a GOP independently of the fact that an atom coefficient decreases significantly its amplitude. While this is done in order to keep the encoding algorithm simple for the present work, it makes difficult to properly represent sequences with occlusion or changes through time.

As described in Sec. II-A, low frequency components are represented separately under the form of a highly downsampled version of the original image. To code this information, low frequency bands are jointly coded by applying a simple temporal wavelet transform on them (spatial wavelet transformation is not possible since the low frequency band is already downscaled as much as possible). Here, a simple Haar temporal transform is applied to each group of low frequency bands, belonging to the same GOP. These are quantized using a dead-zone uniform quantizer and then coded by means of arithmetic coding, following a raster-scan ordering. Dedicated m-ary adaptive statistical context estimation is used for Haar low band and for Haar high subbands.

To guarantee independence among GOPs, all adaptive arithmetic coding contexts are reset at the beginning of every GOP.

2) Influence of Weighted-MP on Compression Results: Curves on Fig. 9 (left) show the gain obtained in terms of rate-distortion (R-D) of the regularized Weighted-MP with respect to the non-regularized MP. A consequence of the regularization turns to be, as expected, the reduction of the amount of necessary bit-rate to represent frame to frame variations. The entropy of the parametric representation gets reduced by the low-pass filtering of parameters imposed by the Weighted-MP selection criteria. Furthermore, this criteria (and motion initialization when no *a priori* is available), reduce the propagation of error in atom parameters, contributing to a better R-D behavior.

Fig. 9 (right) shows the effect of using Weighted-MP compared to MP in terms of rate distortion for Foreman sequence. Both curves show the common drift behavior appearing from the predictive nature of the representation. Notice that the regularized version has a gain between 0.5-1.5 dBs over the non-regularized with 20kbps less. The range of rates appearing in the curves is obtained by exploiting the natural SNR scalability that MP expansions have. For a given bit-rate, video frames are progressively reconstructed by limiting the number of spatial atoms used per frame. In this way, coding costs respect a pre-selected bit-rate.

Coding results on a very geometric sequence are compared to two other coding schemes in order to better understand the interest of studying geometry and motion adaptive video representations. In this case, coding results of the first 64 frames in Foreman QCIF are compared to SPIHT-3D [39] and MP3D [10] (a non-motion adaptive geometric video representation). For this particular sequence full of geometric features, the most comparable scheme in performance is that corresponding to MP3D. At very low bit-rates, average representation of structures using the simpler dictionary of MP3D is more interesting from a R-D point of view. Indeed, at very-low bit-rates, motion information becomes too expensive to code. Signaling atom index from the smaller dictionary in [10] requires less bits than for current Weighted-MP approach where motion needs also to be described. At the same time, the slight picture quality improvement at such coarse approximations does not compensate for the associated increase of information needs. For middle and higher rates, in this example, and despite the temporal drift, the richer dictionary available in our work gives advantage to our decomposition. Indeed, the improved reconstruction quality pays for the increased amount of information required to signal the temporal motion of each of the atoms. Results also show that both geometric approaches are able to exploit spatial correlations better than 3D wavelet approach for this highly geometric sequence. A sufficiently lower number of non-zero coefficients is required in the geometric schemes.

In order to achieve significantly better and consistent results through further sequences, neighboring atoms should be jointly coded. Also an atom tracking scheme free of drift or a proper strategy able to handle, in a robust way, most occlusions and disocclusions should be available. However, at this point, this is beyond the scope of the present work.

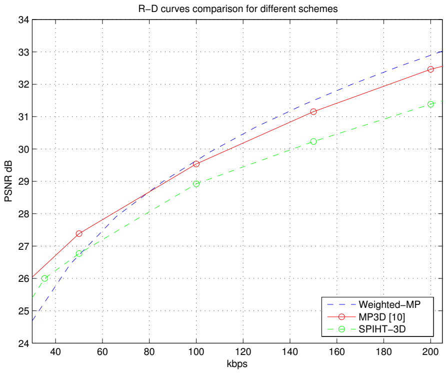


Fig. 10. Comparison of R-D performance for the first 64 frames of the sequence Foreman QCIF using different coding schemes. In particular, current Weighted-MP based scheme is compared to non-motion adaptive MP based MP3D [10] and non-geometry non-motion adaptive wavelet based SPIHT-3D [39].

If *a priori* motion information does not match with the sort of content, then the approximation will not be good. Indeed, spatio-temporal structures are then more difficult to be extracted, and the extracted video features will not capture properly the desired structure. A consequence of this can be seen in Fig. 11. In it we compare the RD performance of 3 representations of the first two GOPs of the sequence Football (QCIF). This part of the sequence contains a series of Football players racing in an erratic motion full of occlusions. As expected, the difficulty of Weighted-MP in capturing the proper signal structure is reflected in a lower R-D performance with respect to the simpler SPIHT-3D. Indeed, the stronger the *a priori* the worse the compression performance. Despite the geometric characteristics of our dictionary, the high drift introduced within the sequence representation makes it difficult to out-perform SPIHT-3D in this case. Further work in proper handling of occlusions and disocclusions is thus required.

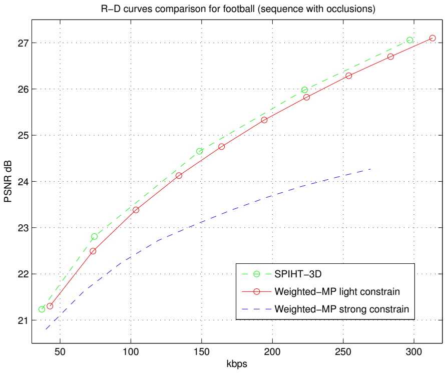


Fig. 11. Comparison of R-D performance for the first 32 frames of the sequence Football QCIF using SPIHT-3D and Weighted-MP with two different *a priori* strengths. Since the *a priori* mismatches with the highly erratic motion in the scene, the higher the λs used, the worse the performance is.

C. Use of Weighted-MP Based Geometric Video Representations for Multi-modal Sequence Analysis : Speaker Localization

Geometry-adaptive video representations can also be used for multi-modal audio-visual analysis. Indeed, the present Weighted-MP based video representation has shown to be an interesting alternative for video feature selection ([40], [41], [42]). These works, however, do not describe in detail the video representation framework used, focusing rather on the aspect of fusing multi-modal data descriptions for audio-visual signal analysis. In [40], [41], [42], a sound is assumed to be generated through the synchronous motion of important visual elements extracted using the proposed video representation framework. Audio and video signals are represented in terms of their most salient structures using redundant dictionaries of functions, making it possible to define acoustic and visual *events*. An audio event is defined as a local maximum (*peak*) of the audio signal energy. A visual event is defined as a peak in the displacement of relevant visual edges represented by video atoms. The intuition behind such a definition is that a maximum in the atoms displacement reflects a movement with respect to a certain equilibrium position, like the one occurring when lips open and close. The synchrony between acoustic and visual events reflects

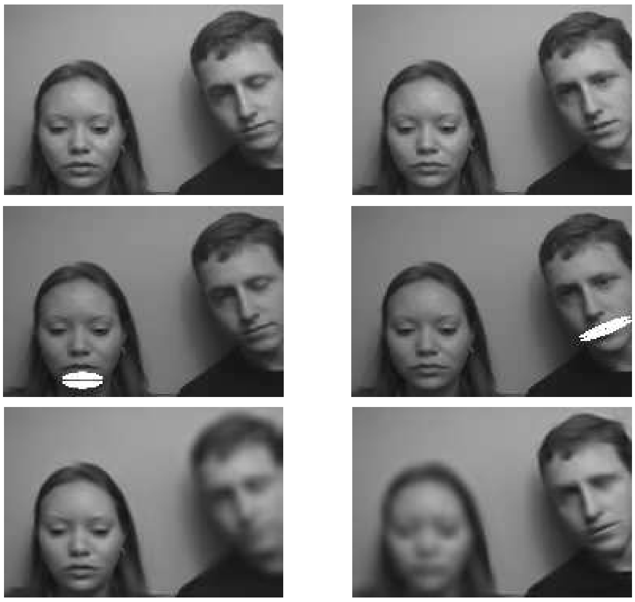


Fig. 12. Sample raw frames (top), video atoms highly correlated with the soundtrack highlighted in white (middle) and reconstruction using only visual structures close to the estimated sound source (bottom). On the first sample the left person is speaking while on the second one the right person is speaking.

the presence of a common source, which is effectively localized. Indeed, the correlation between the audio peaks and the evolution of video atom coefficients is measured in order to determine which are the atoms that are more related to the sound.

Fig. 12 shows an example of the usage of Weighted-MP based geometric video representations for audio-visual data correlation. The test clip involves two persons taking turn in reading series of digits in English. It is recorded at 29.97 fps and at a resolution of 120×176 pixels. The video sequence is decomposed with the proposed algorithm using 200 atoms for the whole scene. To simplify the analysis, no atom refresh mechanism (Sec. IV-D) is used here.

The top row of pictures shows two sample frames taken from the video : on the first one the left person is speaking while on the second one the right person is speaking. In Fig. 12 (middle), the video structures (atoms) that are more correlated with the soundtrack are highlighted in white. In both cases the mouth of the correct speaker is localized. Notice how video atoms nicely adapt their orientation according to the geometric characteristics of the structures they represent.

The Weighted-MP based video representation is very rich and the defined visual structures 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 atoms that are close to the detected sound source. The bottom row of pictures in Fig. 12 shows sample frames of the test sequence reconstructed by summing to the low-pass images only those video atoms that are closer than 80 pixels to the estimated sound source. The reconstructed images can be seen as *audio-visual key frames* that focus on the sound source at a given time instant. Moreover, in a compression application scenario, a sequence may be selectively encoded using only video atoms associated with the soundtrack, saving bits for the coding while keeping the salient information about the scene.

As also shown in [40], [41], [42], results are promising and encourage for further research. The interested reader will find additional results and future development of this research in [43].

VI. Conclusions

In this paper, a study on the representation of video signal as a superposition of spatio-temporal geometric structures has been presented. The purpose of this work is to study the recovery of sparse video approximations where atoms jointly represent spatial geometry and temporal trajectories. In the present work, the problem has been formulated taking into account lessons learned in previous works for the use of redundant coherent dictionaries for sparse approximations. Highly non-linear algorithms such as Matching Pursuit are used together with geometry adapted dictionaries in order to extract geometric primitives from video sequences. The results presented in this work clearly underline that the use of *a priori* information within MP is of key importance for the successful recovery of signal structures with coherent dictionaries. In our investigation, the recently introduced variant of greedy algorithms, Weighted Matching Pursuits [15], is used. This helps to consider motion regularity models for tracking geometric video primitives through time. Results show that Weighted-MP strategies have potential to improve the decomposition of video signals. The obtained representations show to be able to represent and exploit spatio-temporal video structures. Results are promising and encourage for further research in order to better understand the structure of video signals, its proper representation as well as its use in video processing and coding applications. Further research in this direction may include forward-backward extraction of geometric video components, such that occlusions and appearing/disappearing objects are better handled. Also, more accurate *a priori* motion models may be taken into account. In this sense, work is under development in order to study the use of particle filters in the recovery of temporal evolution of video primitives [44]. The use of more signal adapted dictionaries should also be investigated, as well as possible applications of successful representations.

Acknowledgments

The authors would like to deeply thank Dr. Christophe de Vleeschouwer and Prof. Benoît Macq for fruitful discussions and suggestions. Also, many thanks to the reviewers and the editor for their precious feedback that contributed to improve the quality of the paper.

References

[1] O. Divorra Escoda, "Toward sparse and geometry adapted video approximations," Ph.D. dissertation, Ecole Polytéchnique Fédérale de Lausanne, LTS2/ITS EPFL Ecublens CH-1015 Lausanne, Switzerland, July 2005.

[2] E. J. Candès and D. L. Donoho, "Curvelets - a surprisingly effective nonadaptive representation for objects with edges," in *Curve and Surface Fitting*, A.Cohen, C.Rabut, and L.L.Schmaker, Eds. Vanderbilt University Press., 1999.

[3] E. Le Pennec and S. Mallat, "Sparse geometric image representations with bandelets," *IEEE Trans. Image Processing*, vol. 14, no. 4, pp. 423– 438, April 2005.

[4] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, and P. L. Dragotti, "Directionlets: Anisotropic multi-directional representation with separable filtering," *IEEE Trans. Image Processing*, vol. 15, no. 7, pp. 1916–1933, july 2006.

[5] R. M. Figueras i Ventura, P. Vandergheynst, and P. Frossard, "Low rate and scalable image coding using non-linear representations," *IEEE Trans. Image Processing*, vol. 15, no. 3, pp. 726 – 739, 2006.

[6] G. Karlsson and M. Vetterli, "Three dimensional sub-band coding of video," in *ICASSP*, vol. 2. New York: IEEE, April 1988, pp. 1100 – 1103.

[7] S.-J. Choi and J. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Trans. Image Processing*, vol. 8, no. 2, pp. 155 – 167, February 1999.

[8] Y. Andreopoulos, A. Muntanu, J. Barbarien, M. Van der Schaar, J. Cornelis, and P. Schelkens, "In-band motion compensated temporal filtering," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 653–673, Aug. 2004.

[9] P. Frossard, "Robust and multiresolution video delivery: From h.26x to matching pursuit based technologies," Ph.D. dissertation, EPFL, LTS, December 2000.

[10] A. Rahmoune, P. Vandergheynst, and P. Frossard, "MP3D: Highly scalable video coding scheme based on matching pursuit," in *ICASSP*, 2004.

[11] ——, "Flexible motion-adaptive video coding with redundant expansions," LTS-2/ITS - EPFL, Technical Report 17.2004, August 2004.

[12] O. Divorra Escoda and P. Vandergheynst, "Video coding using a deformation compensation algorithm based on adaptive matching pursuit image decompositions," in *ICIP*. Barcelona, Catalonia: IEEE, October 2003.

[13] J. A. Tropp, "Greed is good : Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct 2004.

[14] R. M. Figueras i Ventura, "Sparse image approximation with application to flexible image coding," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Signal Processing Institute, July 2005.

[15] O. Divorra Escoda, L. Granai, and P. Vandergheynst, "On the use of *a priori* information for sparse signal approximations," *IEEE Trans. Signal Processing*, vol. 54, no. 9, pp. 3468–3482, Sep 2006.

[16] P. Vandergheynst and P. Frossard, "Efficient image representation by anisotropic refinement in matching pursuit," in *IEEE ICASSP*, vol. 3, May 2001, pp. 1757–1760.

[17] R. M. Figueras i Ventura, L. Granai, and P. Vandergheynst, "R-D analysis of adaptive edge representations," in *Workshop on Multimedia Signal Processing*, IEEE, Ed., Virgin Islands, December 2002.

[18] R. M. Figueras i Ventura, P. Vandergheynst, P. Frossard, and A. Cavallaro, "Color image scalable coding with matching pursuit," in *ICASSP*, vol. 3, Montreal, 2004, pp. 53–56.

[19] L. Peotta, L. Granai, and P. Vandergheynst, "Very low bit rate image coding using redundant dictionaries," in *Proc. of 48th SPIE Annual Meeting*. San Diego, CA: SPIE, August 2003.

[20] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*, ser. Lecture Notes in Statistics. Springer-Verlag, 1993, vol. 82.

[21] D. L. Donoho, "Wedgelets: Nearly-minimax estimation of edges," *Annals of Statistics*, vol. 27, no. 3, pp. 859–897, 1999.

[22] M. Do, P. Dragotti, R. Shukla, and M. Vetterli, "On the compression of two-dimensional piecewise smooth functions," in *IEEE International Conference on Image Processing (ICIP)*, Thessaloniki, Greece, October 2001.

[23] D. Marr, *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco: W. H. Freeman, 1982.

[24] S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[25] R. M. Figueras i Ventura, O. Divorra Escoda, and P. Vandergheynst, "A matching pursuit full search algorithm for image approximations," ITS-STI/EPFL, Tech. Rep. ITS-2004.031, December 2004.

[26] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.

[27] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[28] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[29] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries-dono," IRISA, Rennes, France, Tech. Rep. 1619, 2004.

[30] V. N. Temlyakov, "Weak greedy algorithms," *Advances in Computational Mathematics*, vol. 12, no. 2-3, pp. 213–227, 2000.

[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., R. L. Howell and J. M. Morriss, Eds. McGrawHill, 1991.

[32] M. P. Queluz, "Multiscale motion estimation and video compression," Ph.D. dissertation, Laboratoire de Telecommunications et Teledetection, UCL, Louvain la Neuve, Belgique, 1996.

[33] T. Aach, A. Kaup, and R. Mester, "Combined displacement estimation and segmentation of stereo image pairs based on gibbs random fields," in *ICASSP*, 1990.

[34] R. Kinderman and J. L. Snell, *Contemporary Mathematics: Markov Random Fields and Their Applications.* American Mathematical Society, 1980.

[35] K. Sayood, *Introduction to Data Compression*, 2nd ed. Academic Press, 2000.

[36] P. G. Howard and J. S. Vitter, "Arithmetic coding for data compression," *Proceedings of the IEEE*, vol. 82, no. 6, pp. 857–865, June 1994.

[37] I. Witten, R. Neal, and J. Cleary, "Arithmetic coding for data compression," *Commun. ACM*, vol. 30, no. 6, pp. 520–540, June 1987.

[38] D. L. Duttweiler and C. Chamzas, "Probability estimation in arithmetic and adaptive huffman entropy coders," *IEEE Trans. Image Processing*, vol. 4, no. 3, pp. 237–246, March 1995.

[39] B. Kim and W. Pearlman, "An embedded wavelet video coder using three-dimensional set partitioning in hierarchical trees (3D-SPIHT)," in *Data Compression Conference*, March 1997.

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

[41] ——, "Analysis of multimodal signals using redundant representations," in *ICIP*, Genova, Italy, 2005.

[42] G. Monaci and P. Vandergheynst, "Audiovisual gestalts," in *CVPR Workshop on Perceptual Organization in Computer Vision*, New York, USA, June 2006.

[43] G. Monaci, "Multimodal analysis using redundant parametric decompositions," on-line: http://lts2www.epfl.ch/∼monaci/multimodal.html, 2005. [Online]. Available: http://lts2www.epfl.ch/∼monaci/multimodal.html

[44] G. Monaci, P. Vandergheynst, E. Maggio, and A. Cavallaro, "Tracking atoms with particles," in *ICASSP*, Hawaii, USA, April 2007.