Perceptual principles for video classification with Slow Feature Analysis

Christian Thériault[(1)], Nicolas Thome[(1)], Matthieu Cord[(1)], Patrick Pérez[(2)]
[(1)]UPMC-Sorbonne Universities, Paris, France [(2)]Technicolor, France
theriaultchristian@gmail.com, nicolas.thome@lip6.fr, matthieu.cord@lip6.fr, patrick.perez@technicolor.com

***Abstract*—At the core of vision research is the notion of perceptual invariance. The question of how the visual system is able to develop stable or invariant states through the ever transforming environment is central to understanding the brain's recognition process. The coined term *slowness principle* used in *slow feature analysis* is a reference to the brain's ability to generate slow changing and thus stable percepts in response to the fast varying visual stimulations in the environment. Based on this principle this paper deals with categorization of video sequences composed of dynamic natural scenes. Unlike models relying on supervised learning or handcrafted descriptors, we represent videos using unsupervised learning of motion features. Our method is based on: 1) Slow feature analysis principle from which motion features representing the principal and more stable motion components of training videos are learned. 2) Integration of the local motion feature into a global classification architecture. Classification experiments produce 11% and 19% improvements compared to state-of-the-art methods on two dynamic natural scenes data sets. A quantitative and qualitative analysis illustrates how the learned slow features untangle the input manifolds and remain stable under various parameters settings.**

I. Introduction

Video understanding can be related to visual perception research which aims at understanding the way in which time and space are integrated by the human visual system. Developing efficient motion descriptors is not only central for video analysis systems, it can also help at gaining a better understanding of perception principles. Motion features usually arise from the relative motion between the different objects in a scene and the eye. Depending on the scene dynamics, motion features range from simple geometrical transforms to more complex non linear transforms all of which do not change the scene identity or, more generally, its category.

This paper addresses the categorization of dynamic natural scenes (*e.g.,* Fire, Rivers, Storms, Lighting, Avalange, *etc.*), see Figures 1, 4. In such type of dynamic scenes, object motion is often correlated with other spatial and temporal variations not intrinsically related to the actual object and overall scene identity: shadows, lighting variations, specular effects, *etc.* Handcrafted descriptors used in the computer vision community, such as HoF or HoG computed on STIP [26], which proved to be very effective for human action recognition, can be quite sensitive to such space-time variations and are thus unlikely to generalize well in this context. Sensitivity to space-time noise or high frequency variations not correlated to objects identity make it difficult to learn internal representations for good classification.

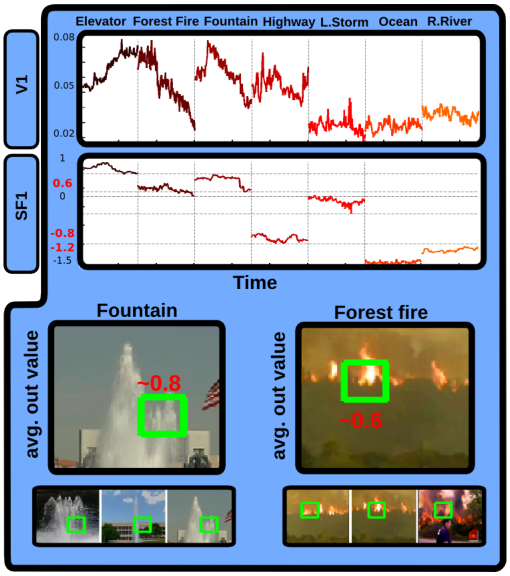


Fig. 1. Top: V1 features generate *tangled up* class representations. However, SF1 (the *slowest* feature learned with SFA) correctly untangles the classes. Bottom: SF1 reveals stable motion components which correlate with semantic categories: upward/backward water motion (fountains/waterfalls), complex flame motion (Forest Fire).

This can also be the case for certain motion features with good neurophysiological inspirations [32], [40] but which remain used in isolation with no further processing or learning of statistics on visual inputs. On the other hand, deep learning of internal representations is an important topic in both A.I. and computational neuroscience [3]. It recently received more attention with its successful application in the context of large scale image classification (Large Scale Visual Recognition Challenge 2012 [ILSVRC2012][3]). Internal representation learning for classification can be referred as the *class manifold untangling* problem [13]: high level representations are expected to be well separated for different semantic categories.

Extending and completing (section IV) previous work in [46], this paper presents an unsupervised method to learn local motion features which self-adapt to the stable temporal components of dynamic natural scenes. For this purpose,

This work was partially supported by ANR ASAP (09-EMER-001).

[3] http://www.image-net.org/challenges/LSVRC/2012/

our model relies on Slow Feature Analysis (SFA) [49]. The principles of SFA point to potential mechanisms by which the brain learns invariant representations: temporal coherence of transforming (i.e. moving) objects may be used by the brain to untangle the factors of variations of visual inputs [13]. The idea that cells in the visual cortex respond to temporal inputs by learning the regularities of our visual world to create stable representations is the basis of several models of cortical processing [22], [49], [48]. Neural activation properties have also been modeled in [5], [4], [25] where the formalism of unsupervised learning of slow varying factors provides a mathematical description for the behavior of complex cell in the visual cortex. SFA learns stable features from a quickly varying input signal. Figure 1 illustrates this principle, by showing how SFA can map representations, which are otherwise not linearly separable, into a space where they can now be *untangled* and classified. On the figure, the mean temporal signal over each class are compared for V1 features[4] and for learned motion features. As shown, the internal signal representation obtained with V1 features is *tangled up* and does not provide sufficient separation between classes for accurate classification. On the other hand, the slowest learned feature (SF1) does provide sufficient separation to correctly untangles the classes: generating outputs with stable responses inside categories and yet different responses between categories. In this simple example, one single slow feature is able to untangle 7 video classes. A larger system works in practice in a larger dimensional space composed by several slow features, allowing even more classes to be correctly separated. At the bottom, figure 1 illustrates how simple slow features learned with SFA can reveal motion features correlated with the semantic classes: upward/backward water motion (fountains/waterfalls), complex flame motion (Forest Fire), *etc*.

The remainder of the paper is organized as follows. Section II overviews related works and positions the paper with respect to more similar approaches. Section III explains the model introduced in this paper: SFA-based learned local motion features and their embedding into a global classification framework. Section IV reports classification scores on two dynamic scenes data sets, pointing out the level of classification achieved. Finally, section V provides a concluding discussion and gives potential directions to be taken from here.

II. Related work & Contributions

This section presents video classification approaches related to this paper. Two main aspects of these related approaches are focused on: 1) the motion features 2) their use for video categorization.

Models of scene classification found in the literature are mostly based on local motion features responding to space-time variations. Although they can be inspired by empirical data (i.e. local visual receptive fields corresponding to directional derivatives) or borrowed from established signal processing approaches, these features are often a design choice optimal for specific applications and are not learned from the statistics of training images. Known examples of such designed motion features are based on optical flow measurements. A specific application is found in [14] where optical flow measurements are used to classify global human actions viewed from a distance using low resolution windows (*i.e.* 30 pixels high). Another use of optical flow applied to natural scenes classification is presented in [26], [27], [30]. Based on Histograms of Optical Flow (HOF) this approach is similar to the static images feature SIFT [29] or HOG [10]. Although theoretically appealing for motion representation, optical flow based approaches are restrained by the optical flow constraints [2], [16], *i.e.* assumes constant illumination between subsequent frames. For this reason, the implementation in practice is not obvious and the performance of this type of motion features is subject to collapse under the context of natural video scenes. For instance, natural space-time variations produced by shadows, lighting variations and specular effects do not change the scenes identity but are intrinsic to natural motions such as fire, waterfalls, river, lighting, avalanges, *etc*. In such type of scenes, far from controlled *in vitro* conditions, direct measurements respecting the optical flow assumption do not follow automatically.

With the aim of explicitly modeling the texture dynamics found in natural environments, linear dynamical systems (LDS) are presented in [42]. This type of stochastic models can be successfully applied in various contexts, from dynamic texture classification to motion segmentation [9] or tracking [8]. However, not unlike optical flow, LDS must respect constraints which are not easily satisfied in complex natural scenes. Therefore, as experimentally reported in [39], due to the first-order Markov property restriction and the linearity assumption, these models might be too restrictive to be applied directly to unconstrained dynamic scenes classification as addressed in this paper.

Biologically inspired features, based on neural processing models of the visual cortex, have been used for object and scene classification tasks [38], [33], [47], [20]. For the case of dynamic scenes classification, some motion features [15], [23] also have explicit biological inspirations. These features can be related to neuro-physiological recordings from the V1-V2-V4 cortical areas which are known to process local spatio-temporal informations [32] and from the MT area which is believed to integrate global motion patterns [40]. This type of spatio-temporal biologically inspired feature has been shown to emerge from the learning of natural images sequence statistics [35].

The problem of dynamic natural scene classification treated in this paper has also been the focus of two recent papers [12], [39]. The work in [12] is based on spatio-temporal filters (i.e. 3d Gabors) energy histograms. The work in [39] focuses on extracting dynamic invariants in chaotic systems. Although both works address the same classification problem as this paper, the approach and method presented here are different, with the focus being on unsupervised motion feature learning.

Minimizing the temporal variations created by motion in order to learn stable representations of objects undergoing

[4] In our approach, each region is described using V1-like features [47], which are effective biologically-inspired image descriptors. The untangling problem illustrated here still holds for various kinds of image features.

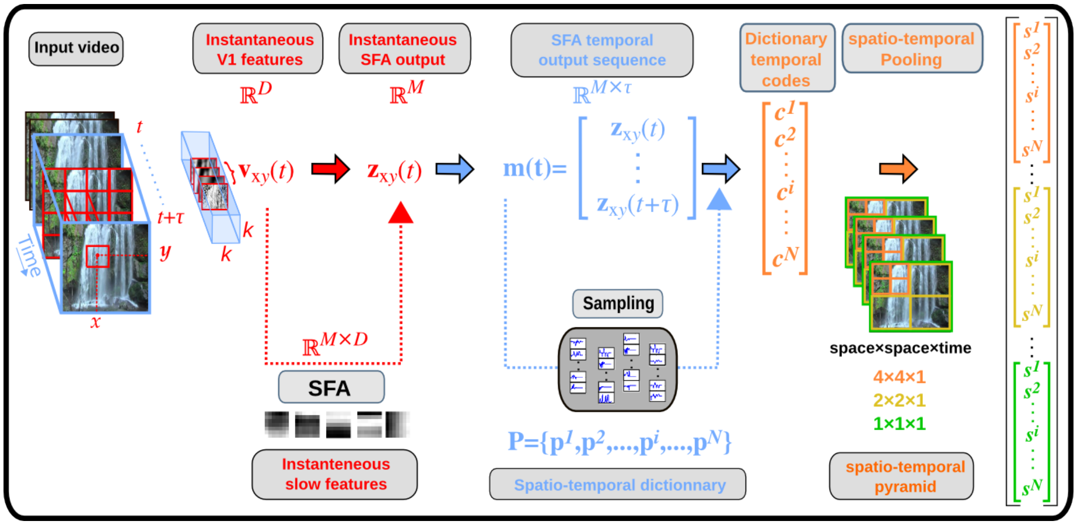

Input video
Instantaneous V1 features
Instantaneous SFA output
SFA temporal output sequence
Dictionary temporal codes
spatio-temporal Pooling
$\mathbb{R}^D$
$\mathbb{R}^M$
$\mathbb{R}^{M\times\tau}$
t
t+τ
y
x
Time
k
k
$v_{xy}(t)$
$z_{xy}(t)$
$m(t)=\begin{bmatrix} z_{xy}(t) \\ \vdots \\ z_{xy}(t+\tau) \end{bmatrix}$
$c^1, c^2, \ldots, c^i, \ldots, c^N$
$\mathbb{R}^{M\times D}$
SFA
Instantaneous slow features
Sampling
$P=\{p^1,p^2,\ldots,p^i,\ldots,p^N\}$
Spatio-temporal dictionnary
space×space×time
4×4×1
2×2×1
1×1×1
spatio-temporal pyramid
$s^1, s^2, \ldots, s^i, \ldots, s^N$

Fig. 2. Dynamic scene classification process. Red: Each video is processed into local regions of V1-like features. These features are then mapped on a set of learned slow features through the SFA principle. Blue: Temporal sequences of slow feature codes are used to train a dictionary of motion features. Orange: Motion features from new videos are mapped on the dictionary before being pooled across time and space into a final vector signature.

transformation is a known principle of unsupervised learning in neuroscience [37], [17], [31], [36], [48]. Although sensory signals (i.e. retinal activation) of a single transforming object vary over a very small time scale, the brain may still use the temporal contiguities (i.e. temporal correlations), intrinsic to this sensory signal to learn stable (slow varying) representations preserving the object's identity [49], [13]. One interesting formalization of this principle, is the *Slow Feature Analysis* model (SFA) [49]. The underlying principle of SFA is that perceptions of objects vary on a slower time scale compared to the input signals from the environment as well as the early processing at the retinal level. In response to temporal input sequences (i.e. motion), the SFA model learns to generate *slower* and thus more invariant output signals. More recently, SFA has been applied and expended in [50] to classify human actions. Closely related to the present paper, this work highlights the relevance of using SFA to extract meaningful motion patterns for video classification.

The SFA principles can be understood from different view points and shown to be equivalent to other signal processing formalisms as well as with synaptic neural learning principles. In [24], SFA is shown to be equivalent under the right conditions, to the Fisher linear discriminant (FLD) method. SFA also shares intrinsic properties with Independent component analysis (ICA) [6] and with the dimension reduction method of Laplacian eigenmaps as demonstrated in [44]. From the point of view of neural coding, SFA can be shown to emerge from learning principle of spike-timing dependent plasticity and even from the Hebbian learning principle (under the correct assumption) [45].

Beyond local representation of motion, a global representation pattern (i.e. vector signature) is usually sought in order to perform classification of scene videos. Beginning with a set of motion features, several possibilities can be considered to create a final global representation. One approach towards a global representation is to directly use global motion descriptors [14], [42] which span the entire spatial area of the scene to be classified. Although appealing, these holistic representations are less robust than systems based on local features. Some models using local motion descriptors are extensions of the BoW framework [41], [1] from static images to video classification [26], [27]. In these models, local motion features (HOF) are extracted at *Space Time Interest Points* (STIP) and mapped on a learned dictionary of features to create a basic code. The coded features can then be pooled across time and space into a final signature used for classification. Certain models with biological inspiration also use this coding and pooling approach [15], [23]. Related to the model presented here, the work in [50] uses the SFA principle to transform videos into histograms of *slow feature* temporal averages. However, the temporal dimension of the input signal is reduced to a scalar value before being accumulated into histograms with no further coding or pooling.

This paper presents a novel method for dynamic scene classification. The entire process is simplified in figure 2. Beginning with a video as input, each frame is processed to extract V1-like features [38]. As a result, each local regions (4×4 in this case) is represented with a vector in $\mathbb{R}^D$ where D is the dimension of V1 space (space, scales and orientations). Each region is then further processed by mapping the V1 features onto a set of $M << D$ slow features, generating a local low dimension representation of size $\mathbb{R}^M$. Mapping high dimensional biological features to a lower dimension manifold for the task of static scenes classification has been explored in [43]. Here, the dimension reduction (SFA) of such biological features is learned and applied on dynamic scenes. The Slow Features Analysis is computed during a learning phase on the entire database of local regions, as explained in section III-A. SFA gives a set of elementary motion patterns as outputs, in a similar manner as done in [50] for human action recognition. However, the present approach differs on many levels. The data set in [50] is concerned with human motion recorded in stable and controlled environments (i.e. uniform background), with very little or no interference. The classification context of this paper is different with the data sets being composed of complex natural scenes. This paper demonstrates that the SFA principle generates a significant untangling of the semantic class manifolds even in the context of complex natural scene videos. Also, the SFA principles are applied on a multi-dimensional V1 representation [38] as opposed to pixels. Furthermore, the temporality of the scene is explicitly maintained all the way into the internal video representation: SFA codes are threaded along τ frames, such that local regions are represented with temporal outputs of size $\mathbb{R}^{M \times \tau}$. Finally, these local spatio-temporal features are embedded inside a global classification architecture, as detailed in section III-B. Here, the difference with respect to [50] is significant since we maintain the full temporal dimension of the input signal into the internal representation. This gives a richer temporal categorical information compared to an averaging method.

In summary, the paper presents the following three main points:

- A local motion descriptor adapted to complex dynamic scenes is introduced. This descriptor is learned in an unsupervised way through the SFA algorithm [49]. SFA generates a low dimensional and low variational subspace representing the stable components of motions across the video frames. Qualitative and quantitative analyses are provided to show that SFA significantly facilitates the untangling of the class manifold.
- The paper presents a coding/pooling architecture in which local temporal outputs are mapped and pooled into a global video signatures. The temporal dimension is maintained into the output signal and therefore, categorical information is not diluted as it is the case when using a temporal averaging over the signal.
- The model generates above state-of-the-art results on two natural scenes data sets: near 11% improvement compared to state-of-the-art methods on the data set recently introduced in [12] and near 19% improvement on the unstable and thus difficult set introduced in [39].

III. Methodology

A. Learning local motion features with SFA

Historically, SFA was introduced as a general principle underlying the brain's ability to learn invariant representations from transformation sequences [18], [49]. The simple and intuitive idea of reducing temporal variations in the input signals has also been studied elsewhere as a synaptic learning rule [37], [17], [31]. Stabilized signal representations gained from SFA in response to transformations has recently been used for human action recognition [50]. From these considerations it naturally follows that the SFA principles would be a good choice to extract stable motion features for dynamic scene classification.

SFA learns instantaneous operators (i.e. operators applied to a single video frame). However, the learned operator must satisfy the constraint of having a low temporal output variation (i.e. operators satisfy a function of multiple frames). Specifically, given a D-dimensional temporal input signal $\mathbf{v}(t) = [v_1(t)v_2(t)...v_D(t)]^T$, the SFA algorithm learns a linear mapping $\mathbf{S}(\mathbf{v}) = [S_1(\mathbf{v}),..,S_M(\mathbf{v})]$ with $\mathbf{S} \in \mathbb{R}^{M \times D}$, such that the new representation $\mathbf{z}(t) = [z_1(t)z_2(t)...z_M(t)]^T$ where each dot product $z_j(t) = S_j(\mathbf{v}(t))$ varies as slowly as possible and still retains relevant information (see figure 3).

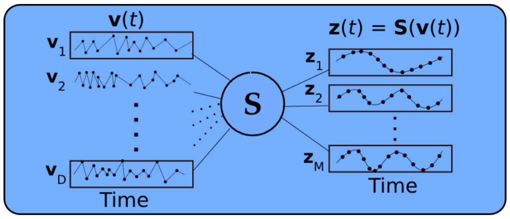


Fig. 3. Slow Feature Analysis. Temporal input signals are transformed into slowly varying signals.

This new representation space is learned by minimizing the average square of the signal temporal derivative

$$\min_{S_j} \langle \dot{y_j}^2 \rangle_t \tag{1}$$

under the constraints:

1) $\langle y_j \rangle_t = 0$ (zero mean)
2) $\langle y_j^2 \rangle_t = 1$ (unit variance)
3) $\forall j < j' : \ \langle y_j, y_{j'} \rangle_t = 0$ (decorrelation)

where $\langle y \rangle_t$ is the temporal average of y.

The above constraints ensure that the output signals vary as slowly as possible while preventing trivial solutions (i.e. constant signals carrying no information). Specifically, the constraints 1. and 2. normalize the outputs to a common scale and prevent the trivial solution $y_j = cst$ associated with a temporal low pass filter (temporal smoothing). These constraints thus guarantee that the slow features S_j are instantaneous and not temporal average functions: the slow features carry time specific information and do not simply average the signals to gain stability. The decorrelation constraint ensures that different slow features carry different informations. The solution to equation 1, with the slow features S_j ranked from the slowest to the fastest, can be obtained by solving the following eigenvalue problem where the slower features are associated with the smaller eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \lambda_M$.

$$\langle \dot{\mathbf{v}} \dot{\mathbf{v}}^T \rangle_t S_j = \lambda_j S_j \tag{2}$$

In [24], the authors establish an equivalence between the optimization done with the Fisher linear discriminant (FLD) and SFA. Starting from a standard supervised learning problem where FLD can be used, a Markov chain is used to convert the training samples into a vectorial sequence where the SFA learning scheme can be applied. Under these assumptions, the FLD in the original space and the SFA in the sequence space solve the same optimization problem. In FLD, the eigenvectors with largest eigenvalues are optimized to maximize the inter-class variance while minimizing the intra-class variance. They correspond to the slowest components of SFA, which are therefore tuned to extract the most significant and stable components of motion (corresponding to inter-class variance) while filtering out less significant temporal variations (corresponding to intra-class variance).

Applied to video frames sequences, the input signal $\mathbf{v}(t)$ can be chosen to be features of many modalities (i.e. colors, gradients, SIFT, HOG). This paper uses biologically inspired *complex cells* V1 features [11], [21] which are known to produce good image representations. These features can be modeled [5] as done in [47] by selecting the local maxima of Gabor filters $g_{\sigma,\theta}$ applied to the input image with orientations $\theta \in \{\theta_1, \theta_2 .., \theta_\Theta\}$ and scales $\sigma \in \{\sigma_1, \sigma_2, .., \sigma_\delta\}$. Specifically, the SFA inputs are local V1 features of size $k \times k \times \Theta \times \delta$ which are flattened into vectors $\mathbf{v} \in \mathbb{R}^D$ as illustrated in figure 2.

The first step in learning slow features from these V1 features is to define the temporal covariance matrix of equation 2. The variables defining this matrix are $\mathcal{N}$ training videos of duration T on a $p \times p$ grid as illustrated in red in figure 2. The V1 feature for video n at spatial position (x, y) and time t is defined by $\mathbf{v}^n_{xy}(t)$[6]. All possible features $\mathbf{v}^n_{xy}(t)$ with temporal derivatives $\dot{\mathbf{v}}^n_{xy}(t)$ are computed . The temporal covariance matrix of equation 2 is then computed by

$$\langle \dot{\mathbf{v}}\dot{\mathbf{v}}^T \rangle_t \; = \frac{1}{p^2 \mathcal{N} T} \sum_{\substack{x=1 \\ y=1}}^{p} \sum_{n=1}^{\mathcal{N}} \sum_{t=1}^{T} \dot{\mathbf{v}}^n_{xy}(t) \dot{\mathbf{v}}^n_{xy}(t)^T \tag{3}$$

The M slowest features $S_1(\mathbf{v}), .., S_M(\mathbf{v}) \in \mathbb{R}^D$ are the eigenvectors of the above matrix associated with the M smallest eigenvalues. Satisfying the constraints in equation 1, the slowest features generate the most stable non trivial output signals. As illustrated in the introductory figure 1, these slow features map into a space where the class manifolds are more easily untangled and are thus excellent candidates to define stable and relevant motion features for classification. The next section explains how these slow features can be used to encode local motions features inside a global architecture producing a signature for each video.

B. Coding and Pooling

Motion features are defined by threading together short temporal sequences of SFA outputs. This new representation space explicitly maintains the temporal dimensions on a one-to-one relation with the input signal. Specifically, a motion feature $\mathbf{m}(\mathbf{t})$ at position (x, y) and across time $\mathbf{t} = [t..t + \tau]$ is defined by a short temporal SFA output sequence from τ consecutive V1 features using the matrix product

$$\mathbf{m}(\mathbf{t}) {=} [\mathbf{z}_{xy}(t)..\ \mathbf{z}_{xy}(t + \tau)] = \mathbf{S}[\mathbf{v}_{xy}(t)..\mathbf{v}_{xy}(t + \tau)] \tag{4}$$

With M slow features we have $\mathbf{S} \in \mathbb{R}^{M \times D}$ and from equation 4 we obtain motion features $\mathbf{m}(\mathbf{t}) \in \mathbb{R}^{M \times \tau}$. As illustrated in figure 2, these motion features $\mathbf{m}(\mathbf{t})$ act as spatio-temporal *atoms* corresponding to the stable motion components inside a small space-time window of dimension $k \times k \times \tau$.

Features in this new representation space, where class manifolds are more easily separable, can now be coded and pooled into a final signature representation for each video. For this, a spatio-temporal dictionary $\mathbf{P} = \{\mathbf{p}^1, \mathbf{p}^2, .., \mathbf{p}^N\} \subset \mathbb{R}^{M \times \tau}$ of motion features is learned using a simple unsupervised sampling procedure [38] in which N motion features are sampled on training videos at random positions and times. Many learning procedure could be applied to generate the dictionary, *e.g.,* K-means or sparse dictionary learning. However, as shown in [7], random sampling can give good performances when used with efficient coding (i.e. soft assignment or sparse coding) [19].

Given a learned dictionary of motion features, a vector signature for each new video is obtained by encoding $\mathbf{m}(\mathbf{t})$ onto $\mathbf{P}$ using the following normalized dot product:

$$c_i = \frac{\mathbf{m}(\mathbf{t})^T \mathbf{p}^i}{||\mathbf{m}(\mathbf{t})|| \cdot ||\mathbf{p}^i||}, \; i \in \{1, .., N\}. \tag{5}$$

Experimentally, normalizing $\mathbf{m}(\mathbf{t})$ and the columns of $\mathbf{P}$ improves performances. Specifically, the temporal codes c_i are computed using soft assignment: mapping each $\mathbf{m}(\mathbf{t})$ on each $\mathbf{p}^i$ at each position (x, y) and time t. A fixed size vector for each video is obtained by pooling the codes c_i inside the subregions of a spatio-temporal pyramid (space×space×time). This spatio-temporal pyramid matching (STPM) extends the SPM pooling principles of [28] to videos. Here, a three level pyramid with partitions $4 \times 4 \times 1$, $2 \times 2 \times 1$ and $1 \times 1 \times 1$ is used, as illustrated in figure 2. The pooling consists of taking the maximum mapping value [38] inside each 21 subregions such that

$$\mathbf{s}^i = \max_{x,y,t} \; c_i(x, y, t) = \max_{x,y,t} \; \frac{\mathbf{m}(\mathbf{t})^T \mathbf{p}^i}{||\mathbf{m}(\mathbf{t})|| \cdot ||\mathbf{p}^i||} \tag{6}$$

After computing vector signatures for all videos, these can be used to feed a classifier (i.e. SVM).

The overall computational complexity of the network corresponds to basic linear algebra operations (i.e. matrix-vector product, eigenvalue decomposition). With a naive convolution algorithm, for an image size of $m \times m$ and V1 filters of size $n \times n$, the complexity at the first level for $\delta \times \Omega$ filters is of order $O(\delta \; \Omega m^2 n^2)$. The complexity of the unsupervised learning of slow features corresponds to a covariance matrix computation of order $O(p^2 \mathcal{N} T D)$ and an eigenvalue decomposition of order $O(D^3)$. Generating the slow feature outputs at the last level corresponds to matrix-vector multiplications applied at all positions over one video (i.e. $O(MDp^2T)$. The final coding over the dictionary of slow features corresponds to dot products of vectors of size $M\tau$ at all positions (i.e. $O(NM\tau p^2 T)$). It is worth noting that the SFA computations are done offline and are not prohibitive for features of reasonable sizes, as it is the case for the type of features used here. The following sections report classification results using the signatures generated by the above architecture.

IV. Quantitative and Qualitative Evaluation

The above model is evaluated on dynamic scene classification using two challenging data sets: the Maryland "in-the-wild" data set [39] and the recently introduced Yupenn Stabilized Dynamic data set [12]. The later is composed of 14 natural scene categories containing 30 videos each with 145 frames on average. The former is composed of 13 natural scene categories containing 10 videos each with 617 frames on average.

[5] Code available at http://webia.lip6.fr/ cord/BioVision/

[6] The V1 features are normalized to a unit sphere [49]

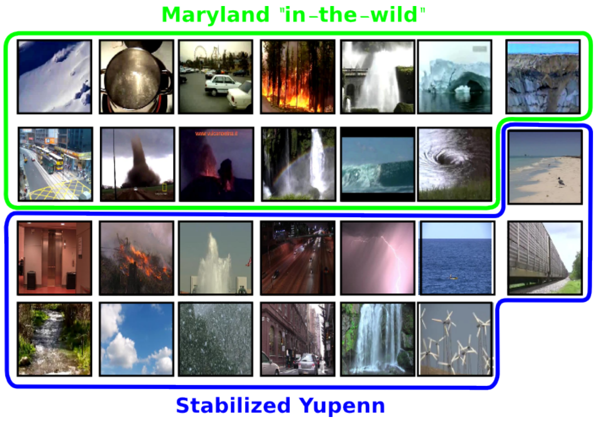


Fig. 4. Samples from Maryland (Top) and Yupenn (Bottom) datasets.

A. *Mono-features results*

First, experimental evaluations of the model using a single visual modality are reported *i.e.* using V1 features as the low-level input features. Accordingly, comparisons are made with state-of-the-art methods based on this mono-feature criterion.

The classification scores reported below are obtained with a linear SVM classifier and computed using the leave-one-out procedure as in [39], [12]. Prior to processing, a basic normalization of video dimensions is applied: all videos are converted to gray scale and resized such that the shortest side has a length of 140 pixels. Basic V1 features are used with $\delta = 2$ and $\Theta = 4$ and $p = 7$ such that $\mathbf{v}_{xy}(t)$ is of dimension $7{\times}7{\times}2{\times}4$. The dimensions of our motion features in equation 4 are set to $M = 30$ and $\tau = 16$ for the Yupenn data set and $M = 100$ and $\tau = 32$ for the Maryland data set. One dictionary element is learned for every 12 training frames on the Yupenn set and every 128 training frames on the Maryland set such that the dictionary sizes are $N = 2440$ and $N = 240$ respectively. As reported below, it is worth mentioning here that classification is quite stable under dictionary size and the above values of N are simply chosen to give slightly better scores.

The detailed results are presented in tables I and II. Although the Maryland set seem to be more difficult (due to unstable camera and drastic view changes), similar conclusions can still be drawn from both data sets. First, the scores obtained are remarkably above all state-of-the-art methods using single modality inputs and well known descriptors. The score of 85.47% (resp. 60.00%) for the Yuppen (resp. Maryland) data set improves by more than 10 pts (resp. 19 pts) the spatial temporal filters (SOE) recently presented in [12], and many other state-of-the-art image and motion features used in computer vision [7]: HoF [30], GIST [34] and Chaotic invariants [39].

A second conclusion from our experiments which is common to both data sets is the improvement gained from our two main contributions, *i.e.* SFA for learning motion descriptors and their embedding in a global representation framework. Column ”No SFA” isolates the effect of SFA learning. The scores obtained when directly using V1 features as input to the coding/pooling architecture, bypassing the SFA representation space, remain relatively high (70.23% and 40.25% in Yupen and Maryland). However, the improvement from learning motion descriptors with SFA is outstanding : $\sim$ 14 pts $\nearrow$ in Yupen, $\sim$ 17 pts $\nearrow$ in Maryland. This validates the relevance of learning motion descriptors which self-adapt to the statistics of training videos and produce a representation space where class manifolds can be untangled. To further isolate the impact of this representation space, we carried out experiments using the SFA embedding proposed in [50]. For an appropriate comparison, we reimplemented their basic method with a similar signature size (i.e. 200 × 21). In both data sets, the performances significantly drops: $\sim$ 66.9% $\searrow$ in Yupen, $\sim$ 33.8% $\searrow$ in Maryland.

	State of the art results				Our re-implementations		
Scenes	HOF [30] [12]	GIST [34] [12]	Chaos [39]	SOE [12]	[50] SFA ASD	No SFA N=2440 $\tau = 16$	Presented model $\tau = 16$ $M = 30$ $N = 2440$
Beach	37	90	27	87	60	70	93.33
Eleva.	83	50	40	67	100	80	96.66
F.Fire	93	53	50	83	46	43	70.00
Fount.	67	50	7	47	46	40	56.66
Highway	30	40	17	77	70	83	93.33
L.Storm	33	47	37	90	70	80	86.66
Ocean	47	57	43	100	83	96	100.0
Rail.	60	93	3	87	76	73	93.33
R.River	83	50	3	93	53	73	86.66
S.Clouds	37	63	33	90	96	93	93.33
Snow	83	90	17	33	63	46	70.00
Street	57	20	17	83	66	93	96.66
W.Fall	60	33	10	43	46	60	73.33
W.mill	53	47	17	57	56	66	86.66
Avg	59	56	20	74	66.9	70.23	**85.47**

TABLE I
CLASSIFICATIONS RESULTS IN AVERAGE ACCURACY FOR THE YUPENN DATA SET. THE MODEL PARAMETERS ARE SET TO $\tau = 16$, $M = 30$, $N = 2440$.

	State of the art results				Our re-implementations		
Scenes	HOF [30] [12]	GIST [34] [12]	Chaos [39]	SOE [12]	[50] SFA ASD	No SFA N=240 $\tau = 32$	Presented model $\tau = 32,$ M=100 N=240
Avalange	0	10	30	10	10	40	60
Boiling.W	40	60	30	60	0	40	70
Chaotic.T	20	70	50	80	80	60	80
Forest.F	0	10	30	40	10	20	10
Fountain	10	30	20	10	20	40	50
Iceberg.C	10	10	10	20	0	30	60
LandSlide	20	20	10	50	20	33	60
Smooth.T	30	40	20	60	30	30	50
Tornado	0	40	60	60	70	80	70
Volcano.E	0	30	70	10	50	60	80
WaterFall	20	50	30	10	30	20	50
Waves	40	80	80	80	50	50	60
Whirlpool	30	40	30	40	70	60	80
Avg	17	38	36	41	33.8	40.25	**60.00**

TABLE II
CLASSIFICATIONS RESULTS IN AVERAGE ACCURACY FOR THE MARYLAND DATA SET. THE MODEL PARAMETERS ARE SET TO $\tau = 32$, $M = 100$, $N = 240$.

Figure 5 gives the confusion matrices for the Yupenn data set [12] and the Maryland data set [39] for our mono-feature model. The classification scores being relatively high, not

[7] re-implemented in [12]

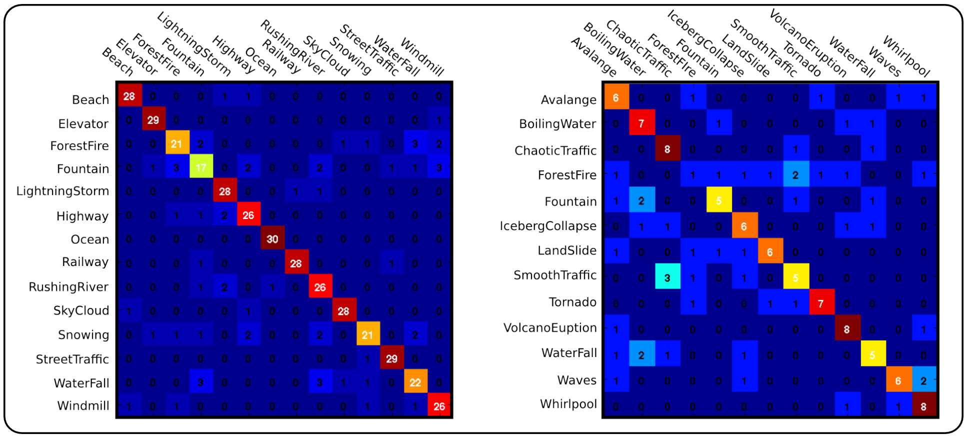


Fig. 5. Confusion matrix for the Yupenn data set (left) and the Maryland data set (right)

many confusion patterns seem to emerge. Some confusions seem to be natural for a system based on motion features without the full context, background knowledge and other perception modalities a human would have.

For example, on the Yupenn data set, two categories which get a relatively high and near symmetrical confusion between them are the water fountain category and the forest fire category. However, these two categories have the lowest overall classification scores. This confusion could still be understood from a pure motion point of view since the motion described by flames and water fountains are similar when discarding color and contextual cues. On the Maryland data set, only the smooth traffic and the chaotic traffic categories seem to display some symmetrical confusion.

Keeping the single modality criterion, the next section illustrates the robustness of the model with respect to parameter variations.

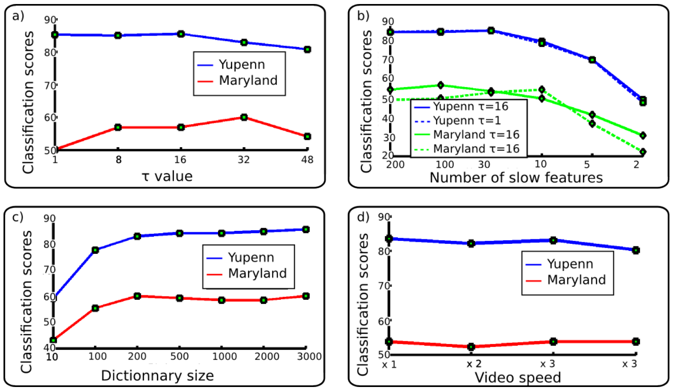


Fig. 6. a) Effect of the temporal length τ of our motion features on classification scores. b) Effect of the number M of slow features on classification scores. c) Effect of the dictionary size on classification scores. d) Effect of video speed on classification scores with a basic setup of $\tau = 8$, $M = 30$ and $N = 192$ (resp.240).

B. *Parameter evaluation*

The effect of the temporal parameter τ, defining the temporal span of motion features, is illustrated in figure 6. On the Maryland set, an increase of up to 10 pts in classification scores is reached when using dictionary elements with a temporal depth of $\tau = 32$. This suggests that more categorical temporal information is captured when using features which span multiple frames, especially on the less stable Maryland data set. Keeping a one-to-one temporal relation between the input and the internal representational space is one major difference with the approach used in [50] in which the slow feature temporal dimension is reduced to a single scalar statistic (i.e. average).

The slowest features carry by definition the most temporally stable components in the scenes (i.e. temporal principal components). As reported in figure 6, classification scores remains relatively high using only a small set of slow features. Being able to keep only the most stable slow features allows for a very compact (i.e. $dim\ \mathbf{z} << dim\ \mathbf{v}$) encoding of the V1 features while still obtaining high classification scores. This makes an efficient representation space.

Another parameter to consider is the dictionary size. Figure 6 shows the effect of dictionary size on classification scores. The scores on both data sets are stable under a wide range of dictionary sizes. High scores are rapidly obtained (before plateauing) using a small dictionary size.

C. *Video speed*

Here, the robustness of the SFA-based technique to the video sampling rate is assessed. For this purpose, the data set is simply down-sampled and classification is performed with the model trained on the initial training videos (without any down-sampling). Figure 6 shows the effect of video speed on classification scores. Videos were artificially speed up by subsampling frames at different rates (2, 3 and 4). Because of the subsampling, many motion artifacts are magnified, and a drop in the results is expected with any basic technique. As shown on the figure, scores remain stable under various video speeds with our SFA-based method. This illustrates the robustness of the SFA *slowness principle* to these temporal distortions. Due to shorter videos in both data sets, videos could only be speed up to a certain speed in order to maintain a sufficient number of frames to fit a least τ frames. Here, a basic setup with $\tau = 8$, $M = 30$ and $N = 192$ (resp.240) is used. The stable classification obtained at different speeds shows that the representation space of slow features remains efficient for classification under different levels of input temporal variations.

D. Further improvement and discussion

1) Multi-features results: To compare the present method with state-of-the art works using multi-features, we combine the above SFA-based framework with other visual modalities. Color and GIST descriptors are used, similarly to what is done in [12]. Thus a simple averaged of RGB values is used as color a descriptor. With a spatial pyramid of size $4 \times 4 \times 1$, as in [12], this leads to a representation for each video of dimension $16 \times 3 = 48$. In addition, GIST descriptors are computed at each frame of the sequence, by computing the energy of a set of Gabor filters applied at different scales (4) and orientations (8) on a spatial (4×4) grid. This leads a representation of size $4 \times 8 \times 16 = 512$ for each frame of the video. As in [39], [12], the mean GIST over the images is used here to represent the whole image sequence, which is thus described by a feature vector of size 512.

The combination presented here is a late fusion scheme: each visual modality is trained separately using a linear SVM, the final decision function f_{comb} is a linear combination of the individual score f_{SFA}, f_{color}, f_{GIST}:

$$f_{comb} = \alpha_{SFA} f_{SFA} + \alpha_{color} f_{color} + \alpha_{GIST} f_{GIST}$$

Two methods to estimate the optimal α_i parameters are used. A first method simply consists in setting them using the individual features performances, as done in [39], [12]. Here we set the α_i in a class-wise manner. We refer to this method as *late fusion baseline*. The second method consists, in cross-validating the α_i parameters. We refer to this method as *late fusion cross-valid*. There are nb classes × nb features parameters (e.g., 13×3 in Maryland), which are uniformly initialized ($\alpha_i = \frac{1}{3} \; \forall i$) and iteratively optimized. To make the cross-validation tractable, we adjust the weights for a given class while the others remain fixed (2 independent parameters).

State-of-the-art results using multi-features are shown in table III. In [39], Chaos + GIST leads to 52% (resp. 58%) with a Nearest-Neighbor (NN) (resp. SVM) classifier. However, in [12], the authors report that they were unable to reach 52% combining only Chaos and GIST and that to achieve such score, color should be additionally incorporated. As indicated in table III, the combination method presented here reaches an improvement of more than 11% for the baseline method.

A simple 2-fold cross-validation on the training set reveals that we can improve performances compared to the late fusion baseline by more than 2 pts, reaching 71.54%. Using more sophisticated cross-validation schemes, one could expect to further improve performances. However, due to the leave-one-out evaluation procedure, cross-validation on the training set is computationally demanding: the parameters must be optimized for each fold of the data (130 times in Maryland). By tuning the parameters by hand, we finally obtained the score of 78.46% (that should not be regarded as state-of-the-art results) illustrating the descriptors complementarity potential.

The results reported in this paper are the same as in [46] for the Yupenn data set but differ for the Maryland data set due to an initial implementation error (code available for download from the authors' website). For the Yupenn data set, the same classification score of $\sim$ 85% as in [46] is obtained with a mono-feature setup (i.e. single modality). The best score obtained using mono-features on Maryland is $60,0\%$, which is 19 pts above the SoA results (41%) at the publication time [12], [39], but below the score reported in [46] ($74,6\%$). However, as shown here, using a multi-feature setup, performances are improved up to a score of 71.5%, which is $\sim$ 14 pts above state-of-the-art results using multi-features (58%, see [39]).

	State of the art results			Presented method	
Scenes	Chaos+ GIST[39] (NN)	Chaos+ GIST [39] (SVM)	Chaos+ GIST+ color[12]	late fusion baseline	late fusion cross-valid
Avalange	40	60	40	50	70
Boiling.W	40	60	40	80	90
Chaotic.T	70	70	70	80	80
Forest.F	40	60	40	90	80
Fountain	70	60	70	80	70
Iceberg.C	40	50	50	50	70
LandSlide	50	30	50	50	80
Smooth.T	50	50	50	50	70
Tornado	90	80	90	90	90
Volcano.E	50	70	50	80	80
WaterFall	10	40	10	50	70
Waves	90	80	90	60	80
Whirlpool	40	50	40	90	90
Avg	52	58	52	69.23	78.46

TABLE III
MULTI-FEATURE CLASSIFICATIONS RESULTS IN AVERAGE ACCURACY FOR THE MARYLAND DATA SET. THE PRESENTED METHOD COMBINES SFA + COLOR + GIST. TWO LATE FUSION METHODS ARE EXPERIMENTED FOR WEIGHTING THE OUTPUT OF THE INDIVIDUAL CLASSIFIERS: LATE FUSION BASELINE AND LATE FUSION CROSS-VALID. SEE TEXT.

2) Temporal split experiment: Here, a different protocol to evaluate the performances on the Maryland set is presented. Each video is split in 3 parts, 2 are used for training and 1 for testing. The classification procedure remains a leave-one-out protocol. However, this time, it's a *leave-one-signature-out* as opposed to a *leave-one-video-out*: for each leave-one-out, different parts (i.e. signatures) of the same video are found in the training and the testing set. The results using a single modality (i.e. V1 features) are reported in Table IV, and show that performances are well increased using this method.

E. Motion feature space

As defined by equation 1, the SFA principle is based on computation of temporal derivatives and therefore assumes a smooth (i.e differentiable) motion pattern. The learned SFA components which map each instant onto a low variation output space are expected to show spatial correlations revealing coherent structures associated with a smooth spatio-temporal pattern. The spatial structure of the learned slow features is illustrated by mapping our motion features into V1 space. Figure 7 displays the V1 projection of the 10 slowest features learned from the Yupenn data set (top) and the Maryland data set (bottom).

As discussed above the SFA learns features which generates slow varying output patterns in response to a fast varying pattern. To illustrates this results, figure 8 shows the smooth

	State of the art results				Our re-implementations	
Scenes	HOF [30], [12]	GIST [34], [12]	Chaos [39]	SOE [12]	[50] SFA Embedding	Our Model Temporal split $\tau = 8$, M=30 N=240
Avalange	0	10	30	10	10	50
Boiling.W	40	60	30	60	0	100
Chaotic.T	20	70	50	80	80	100
Forest.F	0	10	30	40	10	60
Fountain	10	30	20	10	20	90
Iceberg.C	10	10	10	20	0	100
LandSlide	20	20	10	50	20	70
Smooth.T	30	40	20	60	30	70
Tornado	0	40	60	60	70	90
Volcano.E	0	30	70	10	50	80
WaterFall	20	50	30	10	30	90
Waves	40	80	80	80	50	100
Whirlpool	30	40	30	40	70	90
Avg	17	38	36	41	33.8	83.84

TABLE IV
CLASSIFICATIONS RESULTS IN AVERAGE ACCURACY FOR THE MARYLAND DATASET.

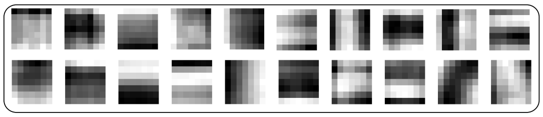


Fig. 7. The 10 slowest features, mapped on V1 space, from the Yupenn data set (top) and the Maryland data set (bottom).

temporal output signal generated by the first slow feature learned on the Yupenn data set in response to a wave pattern.

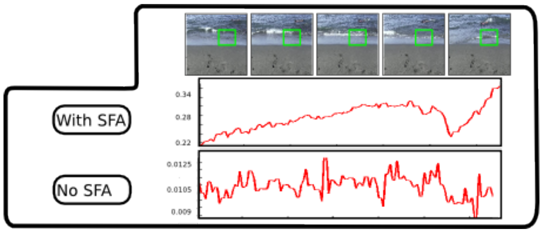


Fig. 8. Temporal output signal of the first slow feature learned on the Yupenn data set (top) and the instantaneous V1 feature (bottom) in response to a wave pattern.

As shown, the output signal of the V1 feature (no SFA) maintains the high variation (i.e. noise) found in the input signal and does not give smooth motion information compared to the slow feature signal (with SFA) which removes high variations not related to the signal's identity. It can also be noted that the SFA signal correlates in time with the motion pattern (the wave), whereas the raw V1 curve has a more random behavior.

As explained in section III, our motion features are defined by the output of M slow features over a duration of τ frames. When using $M = 30$ slow features, the model is able to reach a score of 85.47% on the Yupenn data set. Remarkably, five slow features still reach a score of $\sim$ 73%. This result is already suggested by figure 1 which illustrates the near perfect separation obtained by a single slow feature on 7 classes of the Yupenn data set.

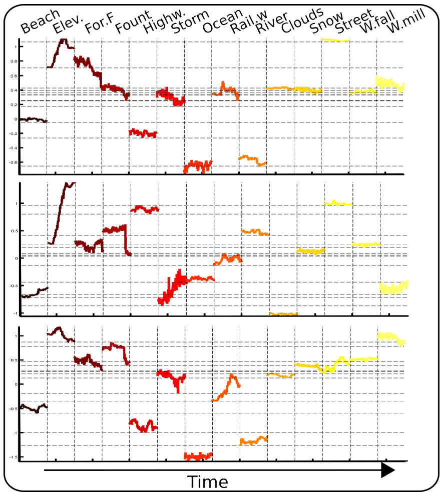


Fig. 9. Semantic untangling of all 14 classes of the Yupenn data set obtained independently by 3 individual dimensions of our motion features. The curves shown (in color to help visualize class separation) are averaged in time over all videos for each category.

Figure 9 illustrates the class manifolds untangling obtained by individual slow features on all 14 classes of the Yupenn data set. A single slow feature cannot untangle all the classes but still achieved an impressive separation using a single dimension. This efficient representation space in which classes are more easily separated generates even better separability when using multiple dimensions (i.e. several slow features) as shown in figure 6.

V. Conclusions and Summary

This paper presented motion features for video scenes classification based on perceptual principles with foundation in neurosciences. These motion features are learned in an unsupervised manner using the neural coding principles of slow feature analysis: they are the result of mapping temporal sequences of instantaneous image features into a low dimensional subspace where temporal variations are minimized. This learned low dimensional representation provides stable descriptions of video scenes which can be used to obtain state-of-the-art classification on two dynamic scenes data sets. One possibility unevaluated in this paper would be to learn stable features from spatio-temporal filters instead of from instantaneous spatial filters. From a general point of view, the classification results reported in this paper suggest the importance of integrating temporal criteria at all stages of recognition, from the input to the internal representation.

References

[1] S. Avila, N. Thome, M. Cord, E. Valle, and A. de Albuquerque. Pooling in image representation: the visual codeword point of view. *CVIU*, 117(5):453–465, 2013. 3

[2] S. S. Beauchemin and J. L. Barron. *The computation of optical flow*. ACM, New York, 1995. 2

[3] Y. Bengio. *Learning Deep Architectures for AI*. Now Publishers Inc., Hanover, MA, USA, 2009. 1

[4] J. Bergstra and Y. Bengio. Slow, decorrelated features for pretraining complex cell-like networks. In *NIPS*, pages 99–107, 2009. 2

[5] P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vision*, 5:579–602, 2005. 2

[6] T. Blaschke, P. Berkes, and L. Wiskott. What is the relation between slow feature analysis and independent component analysis? *Neural computation*, 18(10):2495–2508, 2006. 3

[7] A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011. 5

[8] T. Crivelli, P. Bouthemy, B. Cernuschi-Frias, and J. Yao. Learning mixed-state Markov models for statistical motion texture tracking. In *Proc. ICCV', MLVMA'*, 2009. 2

[9] T. Crivelli, G. Piriou, B. Cernuschi-Frias, P. Bouthemy, and J. Yao. Simultaneous motion detection and background reconstruction with a mixed-state conditional Markov random field. In *ECCV*, 2008. 2

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[11] R. De Valois, E. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22:531–544, 1982. 5

[12] K. G. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *CVPR*, 2012. 2, 4, 5, 6, 8, 9

[13] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73:415–434, 2012. 1, 2, 3

[14] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003. 2, 3

[15] M.-J. Escobar and P. Kornprobst. Action recognition via bio-inspired features: The richness of center-surround interaction. *CVIU*, 116(5):593–605, 2012. 2, 3

[16] D. J. Fleet and Y. Weiss. *Optical Flow Estimation. In Handbook of Mathematical Models in Computer Vision N. Paragios, Y. Chen, and O. Faugeras (eds.).* Springer. 2

[17] P. Foldiak. Learning invariance from transformation sequences. *Neural Comput.*, 3(2):194–200, 1991. 3, 4

[18] M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition with slow feature analysis. In *ICANN*, 2008. 4

[19] H. Goh, N. Thome, M. Cord, and J.-H. Lim. Unsupervised and supervised visual codes with restricted boltzmann machines. In *ECCV*, 2012. 5

[20] Y. Huang, K. Huang, D. Tao, T. Tan, and X. Li. Enhanced biologically inspired model for object recognition. *Trans. Sys. Man Cyber. Part B*, 41(6):1668–1680, 2011. 2

[21] Hubel.D and Wiesel.T. Receptive fields of single neurones in the cat's striate cortex. *J.Physiol*, pages 574–591, 1959. 5

[22] J. Hurri and A. Hyvrinen. Temporal coherence, natural image sequences, and the visual cortex. In *NIPS*, 2002. 2

[23] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007. 2, 3

[24] S. Klampfl and W. Maass. A theoretical basis for emergent pattern discrimination in neural systems through slow feature extraction. *Neural Computation*, 22(12):2979–3035, 2010. 3, 4

[25] K. P. Krding, P. Knig, C. Kayser, and W. Einhuser. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 91:206–212, 2004. 2

[26] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *CVIU*, 108(3):207–229, 2007. 1, 2, 3

[27] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2, 3

[28] P. Lazebnik.S, Schmid.C. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR, 2006. 5

[29] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 2

[30] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *CVPR*, 2009. 2, 6, 9

[31] G. Mitchison. Removing time variation with the anti-hebbian differential synapse.*Neural Comput.*, 3(3):312–320, 1991. 3, 4

[32] J. A. Movshon, I. D. Thompson, and D. J. Tolhurst. Spatial summation in the receptive fields of simple cells in the cats striate cortex. *J. Physiol*, 283:53–77, 1978. 1, 2

[33] J. Mutch and D. G. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *IJCV*, 80(1):45–57, Oct. 2008. 2

[34] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 6, 9

[35] B. A. Olshausen. Learning sparse, overcomplete representations of time-varying natural images. In *ICIP*, 2003. 2

[36] E. T. Rolls and G. Deco. *Computational neuroscience of vision.* Clarendon Press, 2002. 3

[37] E. T. Rolls and T. T. Milward. A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Comput.*, 12:2547–2572, 2000. 3, 4

[38] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE PAMI*, 29:411–426, 2007. 2, 3, 4, 5

[39] N. Shroff, P. K. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *CVPR*, 2010. 2, 4, 5, 6, 8, 9

[40] E. Simoncelli and D. Heeger. A model of neural responses in visual area mt. *Vision Research*, 38:743–761, 1998. 1, 2

[41] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 3

[42] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. In *ICCV*, 2001. 2, 3

[43] D. Song and D. Tao. Biologically inspired feature manifold for scene classification. *IEEE TIP*, 19(1):174–184, 2010. 3

[44] H. Sprekeler. On the relation of slow feature analysis and laplacian eigenmaps. *Neural Comput.*, 23(12):3287–3302, 2011. 3

[45] H. Sprekeler, C. Michaelis, and L. Wiskott. Slowness: An objective for spike-timing-dependent plasticity? *PLoS Computational Biology*, 3(6):e112, 2007. 3

[46] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. *IEEE CVPR*, 2013. 1, 8

[47] C. Thériault, N. Thome, and M. Cord. Extended coding and pooling in the HMAX model. *IEEE TIP*, 22(2):764–777, 2013. 2, 5

[48] G. Wallis and E. T. Rolls. Invariant face and object recognition in the visual system, 1997. 2, 3

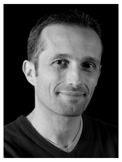
[49] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Comput.*, 14:715–770, 2002. 2, 3, 4, 5

[50] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE, PAMI*, 34(3):436–450, 2012. 3, 4, 6, 7, 9

Christian Thériault received a B.Sc. in mathematics and a Ph.D. in psychology from Université du Québec à Montréal. He also received a B.Sc. from McGill university where he studied physiology and psychology as well as a Master's degree in applied mathematics from université Paris Descartes.

Nicolas Thome received the diplôme d'Ingénieur from the École Nationale Supérieure de Physique de Strasbourg, France, the DEA (MSc) degree from the University of Grenoble, France, in 2004 and, in 2007, the PhD degree in computer science from the University of Lyon, France. In 2008, he was a postdoctoral associate at INRETS in Villeneuve d'Ascq, France. Since 2008 is an assistant professor at Université Pierre et Marie Curie (UPMC) and Laboratoire d'Informatique de Paris 6 (LIP6). His research interests are in the area of Computer Vision and Machine Learning, particularly in the design and learning of complex image representations and similarities, with applications to image and video understanding.

Matthieu Cord received the Ph.D. degree in computer science from the UCP, France, before working at KUL University, Belgium, and in the ETIS lab, France. He joined the Computer Science department LIP6, at UPMC Sorbonne University, Paris, in 2006 as full Professor. In 2009, he was nominated at the IUF (French Research Institute) for a 5 years delegation position. His research interests include Computer Vision, Image Processing, and Pattern Recognition. He developed several systems for content-based image and video retrieval, focusing on interactive learning-based approaches. He is also interested in Machine Learning for multimedia processing, Digital preservation, and Computational cooking. Prof. Cord has published a hundred scientific publications and participated in several international projects (European FP6 and FP7, Singapore, Brazil) on these topics.



Patrick Pérez Patrick Pérez received the Ph.D. degree from the University of Rennes in 1993 and joined INRIA as a Full Time Researcher in 1994. From 2000 to 2004, he was with Microsoft Research Cambridge. In 2009, he joined Technicolor as a Senior Researcher. He is currently a member of the editorial board of the International Journal of Computer Vision. His research interests include image description, search and analysis, and as well as photo and video editing.