Dynamic Scene Classification: Learning Motion Descriptors with Slow Features Analysis

Christian Thériault, Nicolas Thome, Matthieu Cord
UPMC-Sorbonne Universities, Paris, France
theriaultchristian@gmail.com, nicolas.thome@lip6.fr, matthieu.cord@lip6.fr

Abstract

In this paper, we address the challenging problem of categorizing video sequences composed of dynamic natural scenes. Contrarily to previous methods that rely on handcrafted descriptors, we propose here to represent videos using unsupervised learning of motion features. Our method encompasses three main contributions: 1) Based on the Slow Feature Analysis principle, we introduce a learned local motion descriptor which represents the principal and more stable motion components of training videos. 2) We integrate our local motion feature into a global coding/pooling architecture in order to provide an effective signature for each video sequence. 3) We report state of the art classification performances on two challenging natural scenes data sets. In particular, an outstanding improvement of 11% in classification score is reached on a data set introduced in 2012.

1. Introduction

Video understanding has a wide range of application within video indexing, robot navigation and human-computer interaction. Designing efficient motion descriptors is a key ingredient of current video analysis systems. In the most usual context, motion features arise from the relative motion between the different objects in the scene and the camera.

In this paper, we are tackling the problem of categorizing dynamic natural scenes (*e.g.* Fire, Rivers, Storms, Lighting, Avalange, *etc*), see Figure 1. In this context, motion is often correlated with effects that may be considered as interferences or artifacts: shadows, lighting variations, specular effects, *etc*. Therefore, handcrafted descriptors used in the computer vision community, such as HoF or HoG computed on STIP [19], that proved to be very effective for human action recognition, are unlikely to generalize well in our context. The same argument can apply for certain motion features with good neurophysiological inspirations [25, 30] which remain designed and not learned from the statistics of training images. On the other hand, deep learning representation is an important topic in both A.I. and computational neuroscience [3]. It recently received attention with its successful application in the context of large scale image classification, winning the Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)[2]. In the neuroscience community, one challenge of internal representation design or learning is related to the *class manifold untangling* problem [10]: high level representations are expected to be well separated for different semantic categories.

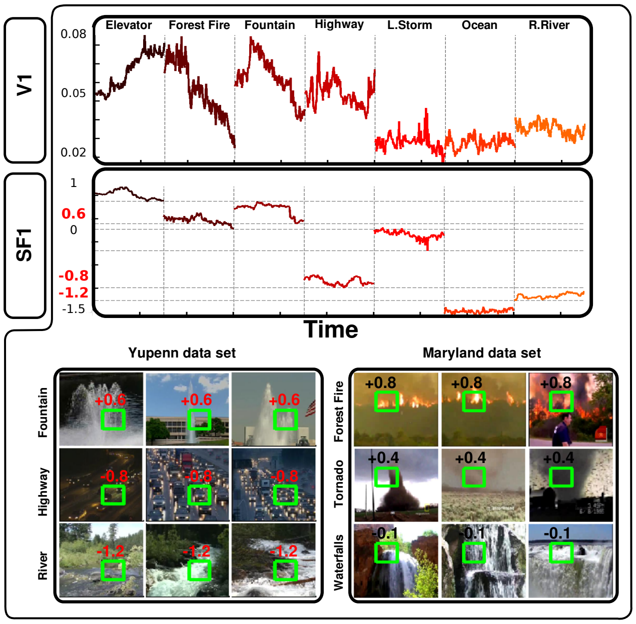

Figure 1. Top: V1 features generate *tangled up* class representations. However, SF1 (the *slowest* feature learned with SFA) correctly untangles the classes. Bottom: SF1 reveals stable motion components which correlate with semantic categories: upward/backward water motion (fountains/waterfalls), complex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

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

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

In this paper, we introduce an unsupervised method to learn local motion features which self-adapt to the difficult context of dynamic scenes. For this purpose, we use the Slow Feature Analysis (SFA) principle which bears foundations in neurosciences [34]. SFA extracts slowly varying features from a quickly varying input signal. Figure 1 illustrates how SFA learning can significantly improve the untangling problem objective. The curves compare the mean temporal signal, over each class, for V1 features[3] and for learned motion features, inside the green windows shown at the bottom. The class representation with V1 features is *tangled up* and cannot separate the classes. On the other hand, the slowest learned feature (SF1) correctly untangles the classes by generating outputs with stable responses inside categories and yet different responses between categories. Quite impressively, one single slow feature is able to untangle 7 video classes. The bottom part of figure 1 illustrates that the slow features learned by SFA reveals sensible motion components 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 2 positions the paper with respect to related works and highlights its main contributions. Section 3 gives the details of the method, introducing our SFA-based learned local motion features and their embedding into a coding/pooling framework. Section 4 reports classification scores on two challenging dynamic scenes data sets, pointing out the remarkable level of performance achieved by the described method using learned motion features. Finally, section 5 concludes the paper and gives directions for future works.

2. Related work & Contributions

In this section, we give more details on video classification approaches related to ours, and focus on two main aspects of the proposed systems: the chosen motion features, and their use for video categorization.

The literature on scene classification includes several handcrafted motion features responding to space-time variations. These motions features are often optimally handcrafted for specific applications and are not learned from the statistics of training images. For instance, in [11], 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 [19, 20, 23]. This motion feature uses Histograms of Optic Flow (HOF) in a similar spirit to the static images features SIFT [22] or HOG [7]. However, because it is restrained by the optical flow constraints [2, 13], *i.e.* assumes constant illumination between subsequent frames, the performance of this type of motion features is subject to collapse under the context of natural video scenes. For example, shadows, lighting variations, specular effects are inherent to motions such as fire, waterfalls, river, lighting, avalanges, *etc*. In this context, the optical flow assumption does not hold. In order to explicitly model texture dynamics, linear dynamical systems (LDS) have been proposed in [32]. Such stochastic models have been successfully applied in various contexts, from dynamic texture classification to motion segmentation [6] or tacking [5]. However, LDS is intrinsically limited by the first-order markov property and linearity assumption. Therefore, as experimentally reported in [29], these models might be too restrictive to properly solve the complex task of unconstrained dynamic scenes classification that we address here. Other motion features [12, 18] presented in the literature are based on 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 [25] and from the MT area which is believed to integrate global motion patterns [30]. These biologically inspired motion features are still not truly learned from stimuli.

Two recent papers have introduced the problem of dynamic natural scene classification [9, 29]. The work in [9] is based on spatio-temporal filters (i.e 3d Gabors), while [29] relies on extracting dynamic invariants in chaotic systems. Although both works address the same classification problem as we do, our approach and method are different as we focus on unsupervised motion feature learning.

One unsupervised learning principle in neuroscience is to minimize temporal variations created by motion in order to learn stable representations of object undergoing motion [27, 14, 24]. One interesting formalization of this principle, is the *Slow Feature Analysis* model (SFA) [34]. The idea behind SFA is that perceptions vary on a slower time scale compared to the input signals from the environment. Given a temporal input sequence (i.e. motion), the SFA model learns to generate a "slower" and thus more invariant outputs signal. Recently, SFA has been investigated in [35] to represent local motion for human action recognition. Interestingly, this work, closely related to ours, consolidates the relevance of using SFA to extract meaningful motion pattern for video classification.

The next step towards classification of scene videos is to obtain one final representation for each video. Given a set of motion features, several possibilities are found in the literature to create a final global representation. One possibility is to use global motion descriptors [11, 32] which cover the entire spatial area of the scene to be classified. However, these holistic representations are less robust than systems based on local features. Other models extend the

[3] In our approach, each region is described using V1-like features [33], which are effective biologically-inspired image descriptors. The untangling problem illustrated here still holds for various kinds of image features.

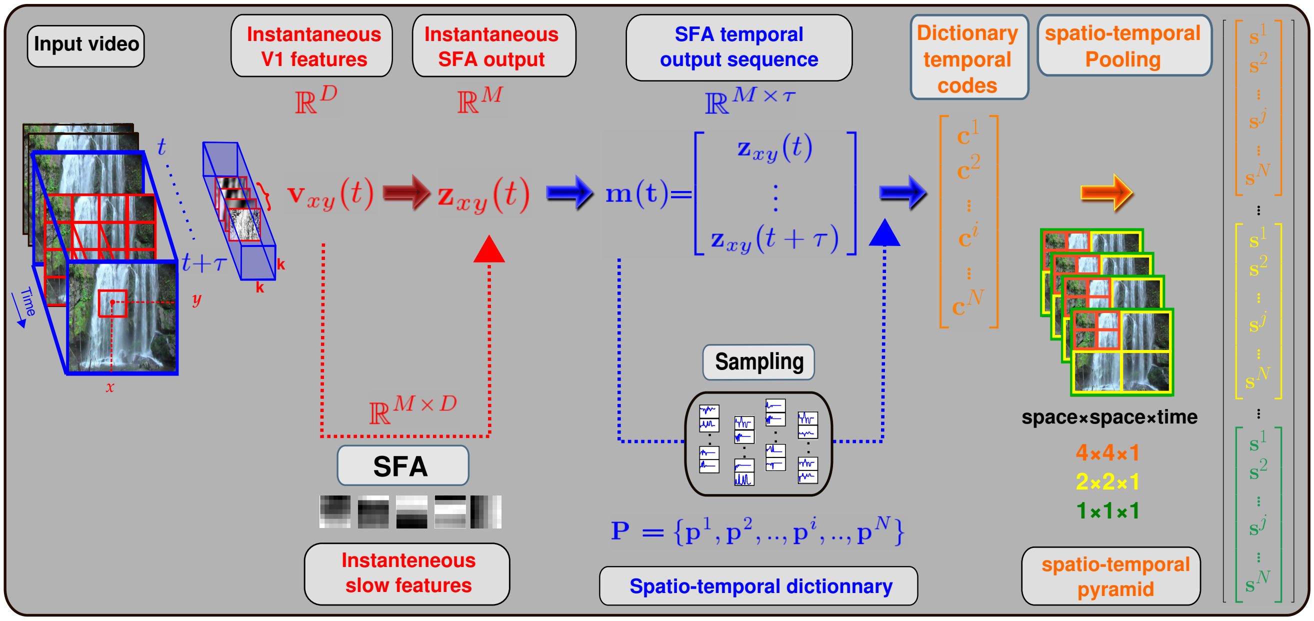
Input video
Instantaneous V1 features
Instantaneous SFA output
SFA temporal output sequence
Dictionary temporal codes
spatio-temporal Pooling
$\mathbb{R}^D$
$\mathbb{R}^M$
$\mathbb{R}^{M\times\tau}$
t
$t+\tau$
Time
y
x
k
k
$\mathbf{v}_{xy}(t)$
$\mathbf{z}_{xy}(t)$
$\mathbf{m}(t) = \begin{bmatrix} \mathbf{z}_{xy}(t)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

Figure 2. Our Dynamic Scene Classification Pipeline. 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.

BoW framework [31, 1] of static images to video classification [19, 20] where local motion features (HOF) are extracted at *Space Time Interest Points* (STIP) and coded by a mapping function on a learned dictionary of features. The coded features can then be pooled into a final signature used for classification. Some models with biological inspiration also use this coding and pooling approach [12, 18]. The work in [35] uses the SFA principle to transforms videos into histograms of slow feature temporal averages. With this approach, the temporal dimension of the input signal is reduced to a scalar value before being accumulated into histograms with no further coding or pooling.

In this paper, we present a novel method for dynamic scene classification. The whole pipeline is depicted in figure 2. For a given video sequence, each frame is processed to extract V1-like features [28], so that local regions (4×4 in our case) are represented with a vector in $\mathbb{R}^D$. Each region is then re-encoded by projecting the V1 features into a set of M Slow Features, leading to a representation of size $\mathbb{R}^M$. The Slow Features Analysis (SFA) is computed offline on the whole database of regions, as explained in section 3.1. SFA ouputs a set of elementary motion pattern, in a similar manner as done in [35] for human action recognition. However, our approach differs at many levels. First, our classification context is different and more challenging. Indeed, their data set is concerned with human motion recorded in stable and controlled environments (i.e. uniform background), with very little or no interference. Second, we show that the SFA principle gives good untangling of semantic class manifolds in the context of complex natural scene videos. Also, we apply the SFA principles on a rich multi-dimensional V1 representation [28] as opposed to pixels. Importantly, to incorporate temporal information in our video representation, SFA codes are threaded along τ frames, so that local regions over time are represented with output sequences of size $\mathbb{R}^{M\times\tau}$. These spatially and temporally localized features are then embedded into a coding/pooling framework, as detailed in section 3.2. Here, the difference with respect to [35] is significant since we maintain the full temporal dimension of the input signal which gives a richer temporal categorial information compared to averaging method. To summarize, the paper presents the three following main contributions:

- We introduce a local motion descriptor adapted to complex dynamic scenes. This feature is learned in an unsupervised way through the SFA algorithm [34]. SFA generates a low dimensional and low variational subspace representing the embedded stable components of motions inside the video frames. We provide qualitative and quantitative analysis supporting the fact that SFA significantly facilitates the class manifold untangling problem.

- We propose a coding/pooling architecture in which temporal outputs sequences of SFA generate global video signatures. By keeping temporal dimension into the output signal, categorial information is not diluted as it is when using a temporal average over the signal [35]. We experimentally report that our embeddeing outperforms the averaging method.

- We report above state-of-the art results on two natural scenes data sets with in particular, 11% improvement compared to state of the art result in the database recently introduced in [9] and near 30% improvement on the challenging set introduced in [29].

3. Methodology

3.1. Learning local motion features with SFA

The SFA principle has been introduced as a mean to learn invariant representations from transformation sequences [15, 34]. The invariance emerging from the SFA principle, which has been used for human action recognition [35], makes it an excellent choice to extract stable motion features for dynamic scene classification.

The SFA principle has the appealing property of learning instantaneous operators (i.e. operators applied to a single video frame) which also satisfy a function of multiple frames (i.e. low temporal output variation). 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 mapping $\mathbf{S}(\mathbf{v}) = [S_1(\mathbf{v}), .., S_M(\mathbf{v})]$ such that the new representation $\mathbf{y}(t) = [y_1(t)y_2(t)...y_M(t)]^T$ where $y_j = S_j(\mathbf{v}(t))$ vary as slow as possible and still retains relevant information (see figure 3).

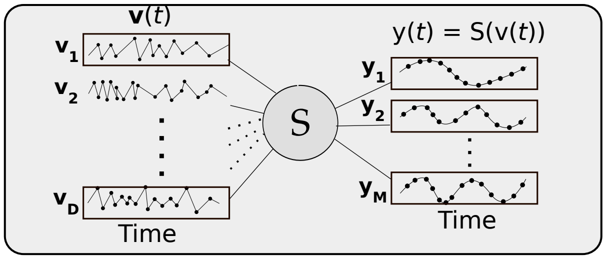

Figure 3. Slow Feature Analysis. Temporal input signals are transformed into slowly varying signals.

This is obtained by minimizing the average square of the signal temporal derivative

$$\min_{S_j} < \dot{y_j}^2 >_t \tag{1}$$

under the constraints:

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

where $<y>_t$ is the temporal average of y. With these constraints the SFA principle ensures that output signals vary as slowly as possible without being a simple constant signal carrying no information. Specifically, constraints 1. and 2. normalize the outputs to a common scale and prevent the trivial solution $y_j = cst$ which would be obtained with a temporal low pass filter (temporal smoothing). Therefore, the slow features S_j must be instantaneous and cannot be averaging the signals over time. This ensures that the slow features carry time specific information and do not simply dilute the signals. Constraint 3. ensures that different slow features carry different informations. The solution to equation 1, with the slow features $S_j(x)$ 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 $\lambda_1 \leq \lambda_2 \leq \ldots \lambda_M$.

$$S_j : \ < \dot{\mathbf{v}}\dot{\mathbf{v}}^T >_t S_j = \lambda_j S_j \tag{2}$$

In our context (video classification), the SFA input signal $\mathbf{v}(t)$ can be image features of many modalities (i.e colors, gradients, SIFT, HOG). We use the biologically inspired *complex cells* V1 features [8, 17] which are known to produce good image representations. These features can be modeled [4] as done in [33] 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 $\sigma \in \{\sigma_1, \sigma_2, .., \sigma_S\}$. Specifically, as illustrated on figure 2, the SFA inputs are local V1 features of size $k \times k \times \Theta \times S$ which we flatten into vectors $\mathbf{v} \in \mathbb{R}^D$ as illustrated in figure 2.

Now, to learn slow features from these V1 features, we need to define the temporal covariance matrix of equation 2. To do this, we consider $\mathcal{N}$ training videos of duration T on a $p \times p$ grid as illustrated in red in figure 2. We define $\mathbf{v}_{xy}^n(t)$[5] as the V1 feature for video n at spatial position (x, y) and time t. We compute all possible features $\mathbf{v}_{xy}^n(t)$ and compute the temporal derivatives $\dot{\mathbf{v}}_{xy}^n(t)$. The temporal covariance matrix of equation 2 is then computed by

$$< \dot{\mathbf{v}}\dot{\mathbf{v}}^T >_t \ = \frac{1}{p^2 \mathcal{N} T} \sum_{\substack{x=1\\y=1}}^{p} \sum_{n=1}^{\mathcal{N}} \sum_{t=1}^{T} \dot{\mathbf{v}}_{xy}^n(t) \dot{\mathbf{v}}_{xy}^n(t)^T \tag{3}$$

The eigenvectors of this matrix associated with the M smallest eigenvalues define our slow features $\mathbf{S}(\mathbf{v}) = [S_1(\mathbf{v}), .., S_M(\mathbf{v})]$. The slowest features generate the most stable non trivial output signals. As previously shown in figure 1, these slow features already produce an impressive untangling of class manifolds and are thus excellent candidates to define stable and relevant motion features for classification. The next section explains how we use these

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

[5] The V1 features are normalized to a unit sphere [34]

slow features to encode local motions features which are then pooled into a final signature for each video.

3.2. Coding and Pooling

Our motion features are defined by threading together short temporal sequences of SFA outputs to generate a new representation space. Specifically, we define a motion feature $\mathbf{m}(\mathbf{t})$ at position (x, y) and across time $\mathbf{t} = [t..t{+}\tau]$ by a short temporal SFA output sequence from τ consecutive V1 features using the matrix product

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

If we use M slow features, then $\mathbf{S} \in \mathbb{R}^{M\times D}$ and equation 4 defines motion features $\mathbf{m}(t) \in \mathbb{R}^{M\times\tau}$. As illustrated in figure 2, these motion features $\mathbf{m}(\mathbf{t})$ can be interpreted as spatio-temporal *atoms* describing the stable motion components inside a small space-time window of dimension $k \times k \times \tau$.

Inside this new representation space, we use a coding/pooling strategy to represent each video by a vector signature of fixed size. In order to do this, we define a spatio-temporal dictionary $\mathbf{P} = \{\mathbf{p}^1, \mathbf{p}^2, .., \mathbf{p}^N\}$ of motion features. We chose a simple unsupervised sampling procedure [28] in which we sample N motion features on training videos at random positions and times. More sophisticated learning procedure could be applied, *e.g.* K-Means or sparse dictionary learning. However, as shown in [4], such random sampling gives very competitive performances when used in conjunction to effective coding schemes such as soft assignment or sparse coding [16].

Once the dictionary is learned, we can compute a vector signature for each new video. To do this, we first encode $\mathbf{m}(\mathbf{t})$ onto $\mathbf{P}$ by computing the following dot product $c_i = \mathbf{m}(\mathbf{t})^T\mathbf{p}^i,\ i \in \{1, .., N\}$. The generated temporal codes c_i are computed by soft assigning each $\mathbf{m}(\mathbf{t})$ on each $\mathbf{p}^i$ at each position (x, y) and time t. To obtain a fixed size vector for each video, the codes c_i are pooled inside the subregions of a spatio-temporal pyramid (space$\times$space$\times$time). This spatio-temporal pyramid matching (STPM) extends the SPM pooling principles of [21] to videos. We use a three level pyramid with partitions $4 \times 4 \times 1$, $2 \times 2 \times 1$ and $1 \times 1 \times 1$ as illustrated in figure 2. The maximum mapping value [28] is pooled inside each 21 subregions such that

$$\mathbf{s}^i = \max_{x,y,t} c_i(x, y, t) = \max_{x,y,t} \mathbf{m}(\mathbf{t})^T\mathbf{p}^i \qquad (5)$$

After pooling, the resulting vector signature can be used to feed a classifier (i.e. SVM). The next section reports classification results using this procedure.

4. Quantitative and Qualitative Evaluation

We evaluate the proposed method for dynamic scene classification in two challenging data sets: the Maryland "in-the-wild" data set [29] and the recently introduced Yupenn Stabilized Dynamic data set [9]. 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. One stunning outcome of our experiments is the large increase in classification performances compared to other state-of-the-art methods.

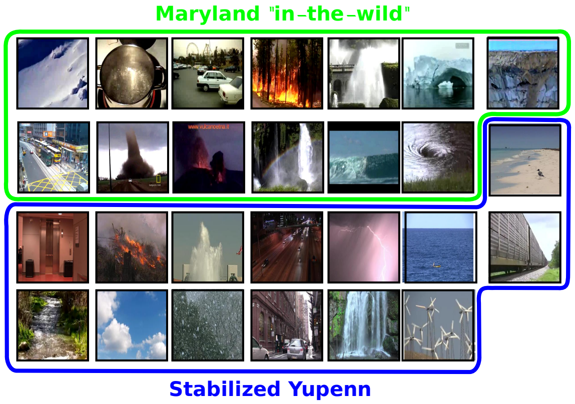

Figure 4. Top: samples from the Maryland "in-the-wild" data set. Bottom: samples from the Yupenn Stabilized data set

4.1. Classification results

All classification scores are obtained with a linear SVM classifier and computed using the leave-one-out procedure as in [29, 9]. All videos are converted to gray scale and resized such that the shortest side has a length of 140 pixels. We used V1 features with $S = 2$ and $\Theta = 4$ and $p = 7$ such that $\mathbf{v}_{xy}(t)$ is of dimension $7 \times 7 \times 2 \times 4$. The dimensions our motion features in equation 4 are setup to $M = 30$ and $\tau = 16$. We learn 1 dictionary element for every 128 training frames such that the dictionary sizes are $N = 192$ and $N = 240$ for the Yupenn data set and the Maryland data set respectively.

The detailed results [6] are presented in tables 1 and 2. Similar conclusions can be drawn from both datasets. First, the scores obtained are remarkably above all state-of-the-art methods using handcrafted descriptors. Our score of 85% (resp. 74.6%) in the Yuppen (resp. Maryland) data set outperforms by more than 10 pt (resp. 30 pt) the recently spatial temporal filters (SOE) proposed in [9], and many other state-of-the art image and motion features used in the computer vision community[7]: HoF [23], GIST [26] and Chaotic invariants [29].

The motivation for using a linear SVM classifier is to highlight the untangling ability of our SFA based representation. To compare our performances with the highest reported score on the Yupenn data set, we also ran classifica-

[6] Confusion matrices can be found in the supplementary material

[7] re-implemented in [9]

tion using the same k-nn classifier as in [9]. We obtained a score of 82% which is close to our SVM score and still well above the results in [9], validating the advantage of our learned representation.

Another experimental result common to both data sets is the favorable impact of our two main contributions, *i.e.* SFA for learning motion descriptors and their embedding in a coding-pooling framework. Column "No SFA" isolates the effect of SFA learning: although the absolute scores remain competitive (57.3% and 55.3% in Yupen and Maryland), the improvement of learning motion descriptors with SFA is outstanding : $\sim$ 30% $\nearrow$ in Yupen, $\sim$ 20% $\nearrow$ in Maryland. This clearly validates the relevance of learning motion descriptors which self-adapt to the statistics of training videos. To isolate the impact of the proposed coding-pooling framework, we carry out experiments by replacing it by the simple embedding proposed in [35]. For a fair comparison, we reimplemented their basic method with a similar signature size (i.e. 200×21). In both databases, the performances significantly drops: $\sim$ 18% $\searrow$ in Yupen, $\sim$ 10% $\searrow$ in Maryland. This illustrates the importance of keeping temporal information and not diluting the signal using a temporal average.

The next section shows how fine-tuning the training parameters can generate even higher classification scores and illustrates the robustness of our results with respect to parameter variations.

	State of the art results				Our re-implementations		
Scenes	HOF [23, 9]	GIST [26, 9]	Chaos [29]	SOE [9]	[35] SFA Embedding	No SFA	Our Model
Beach	37	90	27	87	93	73	96
Eleva.	83	50	40	67	93	46	86
F.Fire	93	53	50	83	76	76	90
Fount.	67	50	7	47	63	56	63
Highway	30	40	17	77	80	33	70
L.Storm	33	47	37	90	53	60	80
Ocean	47	57	43	100	60	76	96
Rail.	60	93	3	87	40	56	83
R.River	83	50	3	93	40	50	83
S.Clouds	37	63	33	90	60	100	100
Snow	83	90	17	33	46	43	73
Street	57	20	17	83	86	26	90
W.Fall	60	33	10	43	70	30	86
W.mill	53	47	17	57	83	73	90
Avg	59	56	20	74	67.6	57.3	**85**

Table 1. Classifications results in average accuracy for the Yupenn Data set

4.2. Parameter evaluation

Figure 5 shows the effect of the temporal parameter τ. An increase of 8% in classification scores is reached when using dictionary elements with a temporal depth of $\tau = 16$. This suggests that more categorial temporal information is captured when using features which span multiple frames. This is one major difference with the approach used in [35] in which the slow feature temporal dimension is reduced to a single scalar statistic (i.e. average).

	State of the art results				Our re-implementations		
Scenes	HOF [23, 9]	GIST [26, 9]	Chaos [29]	SOE [9]	[35] SFA Embedding	No SFA	Our Model
Avalange	0	10	30	10	80	90	90
Boiling.W	40	60	30	60	70	60	80
Chaotic.T	20	70	50	80	50	30	60
Forest.F	0	10	30	40	60	30	80
Fountain	10	30	20	10	60	70	50
Iceberg.C	10	10	10	20	30	40	70
LandSlide	20	20	10	50	70	20	80
Smooth.T	30	40	20	60	50	30	70
Tornado	0	40	60	60	70	50	80
Volcano.E	0	30	70	10	60	60	60
WaterFall	20	50	30	10	60	70	70
Waves	40	80	80	80	100	100	100
Whirlpool	30	40	30	40	80	70	80
Avg	17	38	36	41	64.6	55.3	**74.6**

Table 2. Classifications results in average accuracy for the Maryland Data set

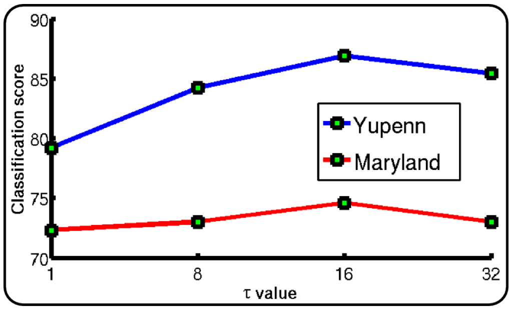

Figure 5. Effect of the temporal length of our motion features on classification scores.

As reported in figure 6, good classification scores are reached using a only a small set of slow features. By keeping only the most stable slow features (i.e. $dim\ \mathbf{y} << dim\ \mathbf{v}$) we obtain a compact encoding of the V1 features and still obtain high classification scores, making this representation very efficient. Another important parameter to consider is the dictionary size. Figure 7 shows the effect of dictionary size on classification scores. The scores on both data sets are stable under a wide range of dictionary sizes, highlighting the robustness of our motion feature representation.

4.3. Motion feature space

The SFA algorithm is based on computation of temporal derivatives and therefore assumes a smooth (i.e differentiable) motion pattern. The stable SFA components are therefore expected to be smooth in both time and space. The smooth spatial structure of learned slow features is illustrated by mapping our motion features into V1 space. Figure 8 displays the V1 projection of the 10 slowest features learned from the Yupenn data set (top) and the Maryland

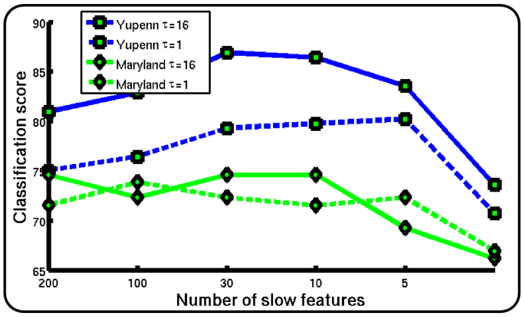

Figure 6. Effect of the number of slow features on classification scores

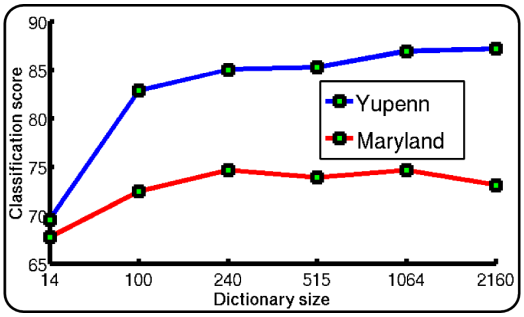

Figure 7. Effect of the dictionary size on classification scores.

data set (bottom). Figure 9 illustrates the smooth temporal output signal from the first slow feature learned on the Yupenn data set in response to a wave pattern. As shown, the output signal of the instantaneous V1 feature (no SFA) does not give smooth motion information compared to the slow feature signal (with SFA). In addition, The SFA signal correlates with semantic motion pattern (the wave), whereas the raw V1 curve has a more random behavior.

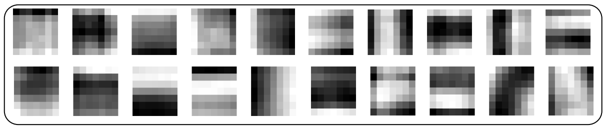

Figure 8. The 10 slowest features, mapped on V1 space, from the Yupenn data set (top) and the Maryland data set (bottom).

As defined in section 3, our motion features are the result of M slow features varying over time. While our full system, using $M = 30$ slow features, reaches a score of 86.9 on the Yupenn data set, one single slow feature (the slowest) still reaches a score of 73.57. This remarkable result is first introduced in figure 1 which illustrates the perfect separation achieved by a single slow feature on 7 classes of the Yupenn data set. Figure 10 complements the results of figure 1 and illustrates the semantic untangling achieved by individual slow features on all 14 classes of the Yupenn data set. As shown, one single slow feature cannot untangle all the classes but still achieved impressive separation using a single dimension. This efficient representantion expresses its full potential when using multiple dimensions (i.e. several slow features) as our classification scores confirms it.

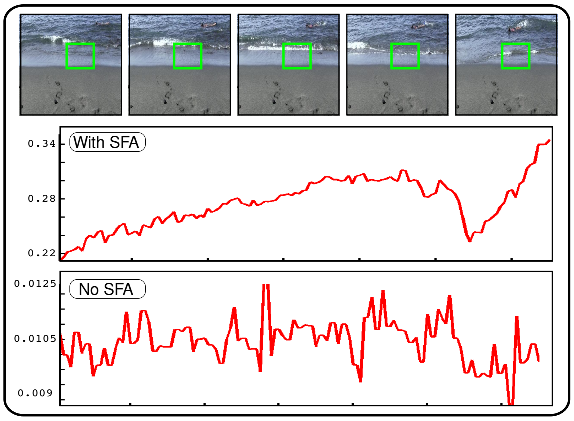

Figure 9. Temporal output signal of the first slow feature learned on the Yupenn data set (top) and the instantaneous V1 feature (bottom) in response to a wave pattern.

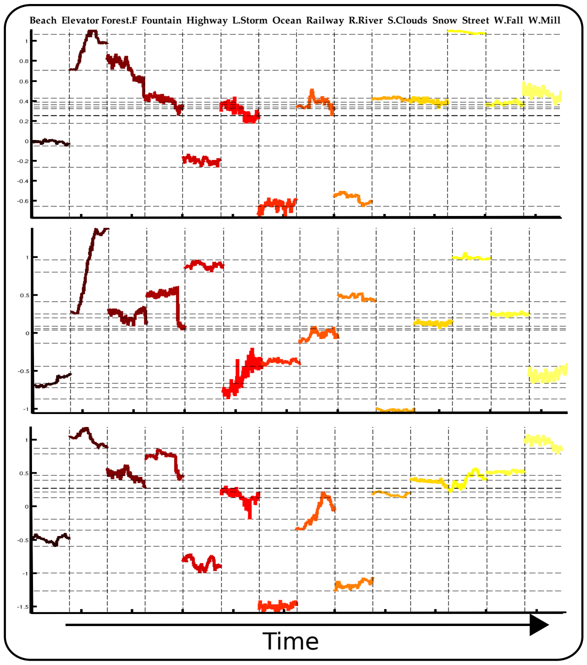

Figure 10. Semantic untangling of all 14 classes of the Yupenn data set achieved independently by three individual dimensions of our motion features. The curves shown are averaged in time over all videos for each category.

5. Conclusions and summary

This paper presented motion features for video scenes classification learned in a unsupervised manner. These motion features 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 challenging 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. The outstanding classification results reported in this paper also suggest that temporal output signals provide more categorical information compared to instantaneous outputs.

As many classes studied in the paper consist of dynamical textures, one interesting direction for future work is to use the classification pipeline for motion segmentation and action recognition.

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, May 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] A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML 2011*, pages 921–928, 2011. 5

[5] T. Crivelli, P. Bouthemy, B. Cernuschi-Frias, and J. Yao. Learning mixed-state markov models for statistical motion texture tracking. In *Proc. ICCV'09, MLVMA'09*, 2009. 2

[6] 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'08*, volume 1, pages 113–126, 2008. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR, vol 2*, pages 886–893, 2005. 2

[8] 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. 4

[9] 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 12*, pages 1306–1313, 2012. 2, 4, 5, 6

[10] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. How does the brain solve visual object recognition? *Neuron*, 73:415–434, 2012 Feb 9 2012. 1

[11] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003. 2

[12] 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

[13] 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

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

[15] M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition with slow feature analysis. In *ICANN*, pages 961–970, 2008. 4

[16] H. Goh, N. Thome, M. Cord, and J.-H. Lim. Unsupervised and supervised visual codes with restricted boltzmann machines. In *ECCV*, pages 298–311, 2012. 5

[17] Hubel.D and Wiesel.T. Receptive fields of single neurones in the cat's striate cortex. *J.Physiol*, pages 574–591, 1959. 4

[18] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, pages 1–8, 2007. 2, 3

[19] 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

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

[21] P. Lazebnik.S, Schmid.C. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. volume 2 of *CVPR*, pages 2169–2178, 2006. 5

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

[23] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE CVPR*, 2009. 2, 5, 6

[24] G. Mitchison. Removing time variation with the anti-hebbian differential synapse.*Neural Comput.*, 3(3):312–320, 1991. 2

[25] 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

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 5, 6

[27] 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. 2

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

[29] N. Shroff, P. K. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *CVPR*, pages 1911–1918, 2010. 2, 4, 5, 6

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

[31] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, 2003. 3

[32] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. In *ICCV*, pages 439–446, 2001. 2

[33] C. Thériault, N. Thome, and M. Cord. Extended coding and pooling in the hmax model. *IEEE Trans.Image processing*, 22(2):764–777, Feb. 2013. 2, 4

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

[35] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE, PAMI*, 34(3):436–450, 2012. 2, 3, 4, 6