Space-Time Robust Video Representation for Action Recognition

Nicolas Ballas
CEA-List & Mines-ParisTech
ballas.n@gmail.com

Yi Yang
Carnegie Mellon University
yi.yang@cs.cmu.edu

Zhen-zhong Lan
Carnegie Mellon University
lanzhzh@cs.cmu.edu

Betrand Delezoide
CEA-List
bertrand.delezoide@cea.fr

Françoise Prêteux
Mines-ParisTech
francoise@mines-paristech.fr

Alex Hauptmann
Carnegie Mellon University
alex@cs.cmu.edu

Abstract

We address the problem of action recognition in unconstrained videos. We propose a novel content driven pooling that leverages space-time context while being robust toward global space-time transformations. Being robust to such transformations is of primary importance in unconstrained videos where the action localizations can drastically shift between frames. Our pooling identifies regions of interest using video structural cues estimated by different saliency functions. To combine the different structural information, we introduce an iterative structure learning algorithm, WSVM (weighted SVM), that determines the optimal saliency layout of an action model through a sparse regularizer. A new optimization method is proposed to solve the WSVM' highly non-smooth objective function. We evaluate our approach on standard action datasets (KTH, UCF50 and HMDB). Most noticeably, the accuracy of our algorithm reaches 51.8% *on the challenging HMDB dataset which outperforms the state-of-the-art of* 7.3% *relatively.*

Figure 1: "Soccer" and "Running" are likely to be distinguished by the area surrounding the human legs while "Clap" and "Wave" are more easily distinguished by the upper-bodies.

Inter-Videos Intra-Video

Figure 2: In different videos, actions localization can be subject to variation due to camera viewpoint change. But, even within a single video sequence, the action area can change among frames.

1. Introduction

With the constant expansion of visual online collections, action recognition has become an important problem in computer vision. It is a difficult task since online videos are subject to large visual diversity. Robust to such variability, Bag-of-Features (BoF) [23] has been adopted as the main paradigm for representing a video. A BoF is computed in 3 steps: (1) local feature extraction, (2) local feature coding and (3) local feature pooling. This paper focuses on the third step that aims at summarizing the feature code distribution in a fixed length vector. Traditional pooling considers each local feature independently [11]. Such an algorithm discards the local feature position information in the video space-volume. However, this space-time context has been proven useful for classification [11]. Indeed, discriminative information is not equally distributed in the video space-time domain as shown by Figure 1. To benefit from this context, spatial pooling [12, 11] divides a video using fixed segmentation grids and pools the features locally in each grid cell. Despite the performance improvement, spatial pooling loses the BoF space-time invariance. Different action instances with various localizations in the space-time volume can result in divergent representations. This problem is severe for the actions which have dramatic space-time variance as illustrated in Figure 2. In this case, spatial pooling divides one action across different grid cells which may lead to a significant performance drop. A BoF representation robust to space-time variance is therefore critical

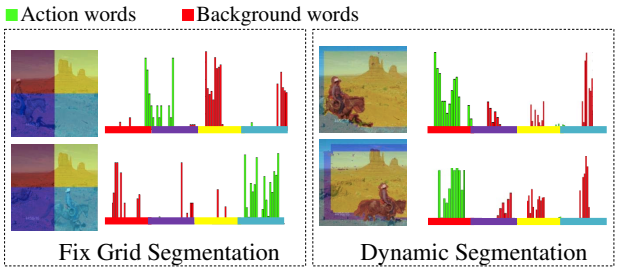


Figure 3: Illustration of the space-time robustness importance.

for action recognition.

In this work, we propose to take advantage of the space-time discriminative context with an emphasis on retaining the space-time robustness. Beyond standard spatial pooling which uses fixed segmentation grids, we segment a video according to its content through saliency maps. Our algorithm relies on the idea that the discriminative information has a non-uniform distribution in saliency spaces. For example, "Running" is more likely to be distinguished from "Walking" by regions subject to high motion. In addition, different saliencies can highlight different regions in the video space-time volumes. They may capture complementary information which can be appropriately fused. Based on those observations, we propose two main contributions.

We introduce a novel space-time invariant pooling which leverages the space-time context. We first extract video structural cues using various saliency measures. We then aggregate the local feature statistics over fixed saliency sub-regions, each sub-region defining a *structural primitive*. Focusing on different structural aspects, *cornerness*, *light* and *motion* saliencies are investigated. *Cornerness* highlights regions repeatable under geometric transformations, *motion* identifies regions with strong dynamics and *light* provides coarse object segmentation.

To automatically determine the optimal *structural primitives* combination associated to a specific action, we introduce a sparse feature weighting regularizer, which is able to assign optimal weights to different feature groups. We integrate the $\|.\|_{2,p}$ norm to a linear SVM classifier and propose a Weighted SVM (WSVM) for action recognition. The WSVM objective function being non-smooth, we propose a new efficient optimization algorithm to minimize it.

2. Related Work

Spatial pooling [12, 11] has successfully demonstrated a performance improvement over classic BoF. However, to be fully effective, feature space-time statistics must align with the segmentation grids due to their fixed aspect ratio. Recent efforts [22, 5, 7, 2] have tried to exploit richer spatial or temporal information by learning segmentation grids adapted to specific task. Jia [7] relies on sparsity to select segmentation grids in an overcomplete basis while Sharma and Harada [22, 5] learn weights scheme associated to predefined segmentation grids. Since all those approaches partition local features in the spatial domain, they are not robust to space-time change. They remain sensible to the action localization variance. In video, Cao [2] proposes a scene-adapted pooling. His approach focuses on modeling only the temporal context. It is also not robust to time variation since the local features are pooled in the temporal domain.

Saliency has already been used successfully in image analysis [14, 15, 16, 21, 17, 26]. Rahtu [17] uses saliency to segment object from image. Wang [26] uses saliency to compute highly discriminative local descriptor. In an image recognition context, Parikhn, Shabaz and Moosman [14, 15, 16, 21] define sparse sampling strategies to detect local features. Our motivation significantly differs from those approaches. We do not use saliency information to sample features but to pool them. We identify prominent regions in a video through saliency to model the space-time context while preserving the space-time robustness.

In the remainder of this paper, we start by introducing our space-time invariant pooling. We then present our WSVM. An evaluation of our proposal is finally performed.

3. Space-Time Robust Representation

Figure 3 compares two pooling schemes using 2×2 static grid segmentation or a dynamic segmentation based on motion saliency. Due to its localization variance, the action falls in different cells of the static grids leading to two spatial BoFs having low-similarity despite depicting the same action. By segmenting the video dynamically, the second pooling scheme remains robust to the action space-time variance while still taking advantage of the local feature space-time context. This motivate us to propose a novel pooling algorithm using video content information.

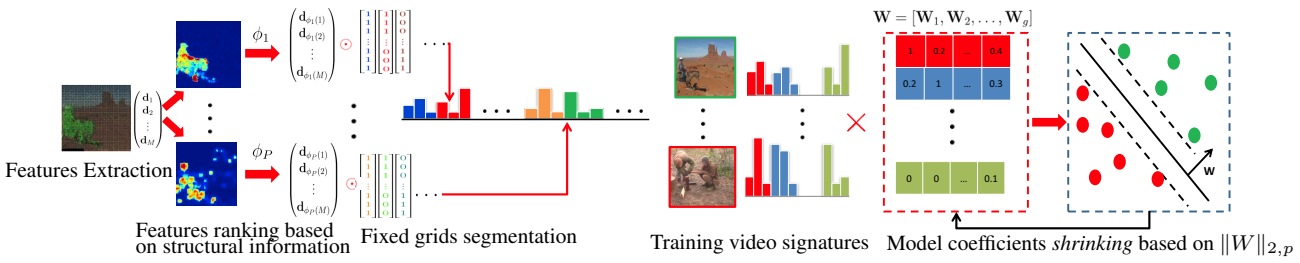
3.1. Content Driven Pooling

In the following, we first reformulate the spatial pooling problem and then extend this formulation to take advantage of video content information.

Let $\mathbf{D} = \{\mathbf{d}_1, ..., \mathbf{d}_M\}$ be a set of local features extracted from a video. We denote by $\mathbf{G} = \{\mathbf{G}^1, ..., \mathbf{G}^n\}$ a set of grid cells. Each $\mathbf{G}^i$ is a binary matrix indicating which video voxels are active, $\mathbf{G}^i \in \{0,1\}^{s_x \times s_y \times s_t}$, (s_x, s_y, s_t) being the video dimension. Based on those definitions, we express the max spatial pooling operation as (1).

$$\mathbf{X}_i = \max_{(x,y,t)} \mathbf{G}^i_{x,j,t} \times code(\mathbf{d}_{\omega(x,y,t)}) \qquad (1)$$

$\omega : \mathbb{R}^3 \rightarrow [1, M]$ is function indexing the descriptors $\mathbf{D}$ based on their positions. The function $code : \mathbf{D} \rightarrow \mathbb{R}^K$ is a local feature coding scheme such as sparse-coding or



(a) Space-Time Invariant Pooling (b) Weighted SVM

Figure 4: Illustration of the space-time invariant pooling and the WSVM algorithm.

locality coding. (1) relies on max pooling since it improves the class separation [1].

Traditional spatial pooling uses a set of pre-defined pyramidal grids segmenting the video in increasingly finer cells. Recent pooling works [22, 5, 7] learn $\mathbf{G}$ directly from data achieving task-specific segmentation. Both approaches pool local features in the space-time domain.

Differently, we aim at modeling the space-time context while remaining robust to the space-time variance. To do so, we identify prominent regions using saliency. As shown in Figure 4a, we (i) extract saliency information from a video, then, (ii) order local features in rank lists according to each saliency and (iii) capture local feature statistics in various rank list sub-regions. As a result, our pooling scheme does not require space-time information to compute video regions, and, it performs video-specific segmentation based on their structural cues. Since our pooling uses ranks to group features instead of absolute values, it remains invariant to global translation in the saliency space.

To formulate our content driven pooling, we modify the indexing function ω in (1) to include video structural cues. Let $\mathbf{P} = \{p_1, ..., p_M\}$ be the saliency values for each local feature. $\phi : [1, M] \rightarrow [1, M]$ is a ranking function ordering the local features according to $\mathbf{P}$. To infer $\Phi = \{\phi(1), \ldots, \phi(M)\}$, we minimize the functional $\min_\Phi \sum_{i=1}^{M} i p_{\phi(i)}$, $d_{\phi(1)}$ is the local features having the highest saliency while $d_{\phi(M)}$ correspond to the lowest one. Our content-based pooling becomes:

$$\mathbf{X}_{i,k} = \max_{j \in [1,M]} \mathbf{G}_j^{i,k} \times code(\mathbf{d}_{\phi(j)}) \qquad (2)$$

With (2), the pooling is performed in the saliency instead of the space-time domain. $\mathbf{G}$ is defined as a pyramidal tiling. We consider sequence of segmentation grids $\mathbf{S}^0 \ldots \mathbf{S}^{L-1}$ such as each grid $\mathbf{S}^i$ is composed by 2^i equally sized cells: $\mathbf{G} = \{\mathbf{G}^{i,1}, ..., \mathbf{G}^{i,2^i}\}$ where $\mathbf{G}^{i,k} \in \{0, 1\}^M$. $\mathbf{G}^{i,k}$ coefficients are equal to 0 except on the interval $G^{i,k}_{[s,e]}$ where $s = \frac{k-1}{2^i} M$ and $e = \frac{k}{2^i} M$.

$\mathbf{X}_{i,k}$ captures the distribution of local features over a saliency sub-region. It defines a *structural primitive*. The *structural primitives* are $\|.\|_2$ normalized and concatenated to obtain the signature $\mathbf{X} = [\mathbf{X}_{1,1}, ..., \mathbf{X}_{L-1,2^{L-1}}]$. When using several saliency functions, we repeat this pooling operation for each measure and concatenate all the resulting *structural primitives*.

3.2. Saliency Measures

To complete the definition our pooling, we need to define the values $\mathbf{P} = \{p_1, ..., p_M\}$. We take advantage of the video visual data through saliency measures to identify prominent or salient areas: $\mathbf{p}_i = s(\mathbf{d}_i)$. $s : \mathbf{D} \rightarrow [0 - 1]$ is a local measure that describes how much a feature differs relatively to its immediate neighborhoods [6]. We focus on 3 different saliency functions : "cornerness", "light" and "motion". The cornerness saliency highlights visually distinctive features, which are repeatable under geometric transformation. Feature cornerness is estimated with the Harris-Laplace transform [14]. The light provides coarse object segmentation. A RGB frame is transformed into the Lab color space. The L (Light) component of the color space is divided in 60 equal-sized bins and the light saliency is computed by an efficient center-surround operation using sliding windows [17]. Motion saliency considers the video optical flow computed for each video frame through the Farneback algorithm [3]. Flow magnitude is quantized into 16 uniform bins. The motion saliency is then computed with the same sliding windows approach as the light saliency [17].

4. Weighting Structural Primitives

As shown in Figure 5, the saliency measures emphasize different areas of the video space-time volume. The discriminative power of those regions is non-uniform and tend to be action dependent, i.e. the saliency measures are not equally discriminative for the different actions. For instance, motion saliency emphasizes foreground as well

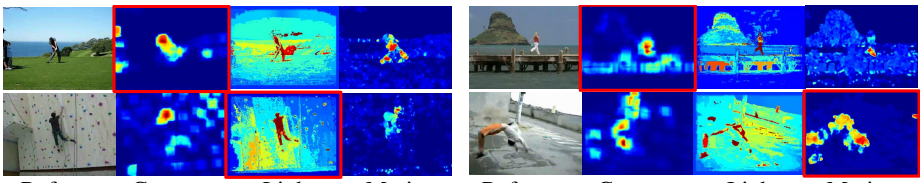


Figure 5: Illustration of prominent areas detected with the different saliency measures. The most discriminative saliency measure for each action is indicated by the red contour.

as background area for an action subject to strong camera movement while light saliency remains robust to this phenomena. By focusing on only a few *structural primitives* at classification, we could take advantage of saliency functions which fit best the action of interest while discarding area containing irrelevant or noisy information.

In this section, we introduce SVM algorithm with a sparse feature weigthing regularizer, illustrated in Figure 4b, that determines the optimal *structural primitives* layout given an action.

4.1. Weighted SVM Model

Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be N training video signatures and $\mathbf{Y} \in \{0,1\}^N$ their corresponding binary labels. Each video signature $\mathbf{X}_i$ is the concatenation of the *structural primitives* i.e., $\mathbf{X}_i = [\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,G}]$. Linear SVM combined to max-pooling has demonstrated encouraging results in the context of image classification while limiting the training complexity to $O(n)$ [27]. We consider a linear model $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_G] \in \mathbb{R}^d$ with its bias term $b \in \mathbb{R}$. $\mathbf{W}_g$ is the group of $\mathbf{W}$ coefficients correlating with the *structural primitive* $\mathbf{X}_{i,g}$. A linear SVM *primal learning formulation* has the following form:

$$E(\mathbf{W}, b) = \sum_{i=1}^{N} L(\mathbf{Y}_i, \mathbf{X}_i\mathbf{W} + b) + \lambda\Omega(W) \tag{3}$$

$$L(\mathbf{Y}_i, \mathbf{X}_i\mathbf{W} + b) = \max(0, \mathbf{Y}_i(\mathbf{X}_i\mathbf{W} + b))^2. \tag{4}$$

L is the square hinge loss and Ω is the regularizing term. The SVM model uses a $\|.\|_2$ norm as regularizer [27]. This norm attaches the same importance to each coefficient in $\mathbf{W}$, i.e., each group $\mathbf{W}_g$ contributes equally. To leverage the non-uniform discriminative power of *structural primitives*, we propose to prioritize only the most substantial groups $\mathbf{W}_g$ for an action while discarding the irrelevant one by adding a sparsity constraint on $\mathbf{W}$.

Sparsity is induced through the use of a $\|.\|_p$ norm with $p < 2$. This method implicitly assumes that each individual coefficient in $\mathbf{W}$ is independent of the all others. It only guarantees sparsity at the $\mathbf{W}$ individual coefficient level and does not assure that a few groups $\mathbf{W}_g$ are prioritized for an action. Group sparsity, on the other hand, uses a $\|.\|_{2,p}$ norm, a combination of a $\|.\|_p$ norm at the groups level and a $\|.\|_2$ norm at the individual coefficient level. While selecting only a few groups with the $\|.\|_p$ norm, it considers the coefficient inside a group as a whole through the $\|.\|_2$, taking advantage of their implicit relation. Hence, a $\|.\|_{2,p}$ regularization term is used our learning formulation in (4), shrinking the number of groups selected.

Algorithm 1 Weighted SVM learning

Input: Input data $\mathbf{X} \in \mathbb{R}^{N\times d}$ and labels $\mathbf{Y} \in \{0,1\}^N$. Regularization parameters λ, p

Output: $\mathbf{W} \in \mathbb{R}^d, b \in \mathbb{R}$

1: Initialize $\mathbf{W} \in \mathbb{R}^d$ and b at random;
2: **repeat**
3: $\mathbf{D} = \begin{pmatrix} (\frac{2}{p}\|\mathbf{W}_1\|_2^{2-p})\mathbf{I}_1 & & \\ & \ddots & \\ & & (\frac{2}{p}\|\mathbf{W}_G\|_2^{2-p})\mathbf{I}_G \end{pmatrix}$
4: $[\mathbf{W}, b] \leftarrow$ L-BFGS$(E, \frac{\partial E}{\partial \mathbf{W}}, \frac{\partial E}{\partial b})$;
5: **until** Convergence

$$E(\mathbf{W}, b) = \sum_{i=1}^{N} L(\mathbf{Y}_i, \mathbf{X}_i\mathbf{W} + b) + \lambda||\mathbf{W}||_{2,p} \tag{5}$$

p controls the group selection sparsity. The smaller p is, the fewer groups are selected by the WSVM. If $p = 2$, we obtain a classic $\|.\|_2$ regularizer term. In this sense, our WSVM model (5) generalizes the classic SVM (4).

4.2. Optimization

We want to minimize with respect to $\mathbf{W}$ and b, our objective function E:

$$\arg\min_{\mathbf{W},b} \sum_i L(\mathbf{Y}_i, \mathbf{X}_i\mathbf{W} + b) + \lambda(\sum_{g=1}^{G} \|\mathbf{W}_g\|_2^p)^{\frac{1}{p}} \tag{6}$$

Due to the $\|.\|_{2,p}$ regularizer, (6) is a non smooth optimization problem. To transform this problem, we rewrite it as (7).

$$\arg\min_{\mathbf{W},b} \sum_i L(\mathbf{Y}_i, \mathbf{X}_i\mathbf{W} + b) + \lambda \sum_{g=1}^{G} \frac{\|\mathbf{W}_g\|_2^2}{\frac{2}{p}\|\mathbf{W}_g\|_2^{2-p}} \tag{7}$$

We define the diagonal block matrix $\mathbf{D}$ (see algorithm 1)[1]:

[1]In practice we add a small ϵ to each diagonal coefficient of $\mathbf{D}$ for numerical stability.

$\mathbf{D}$ is a semi-definite positive matrix. $\mathbf{I}_g$ is the identity matrix corresponding to the group $\mathbf{W}_g$. We deduce that $(\sum_{g=1}^{G} \|\mathbf{W}_g\|_2^p)^{\frac{1}{p}} = tr(\mathbf{W}^T\mathbf{D}^{-1}\mathbf{W}) = \|\mathbf{U}^T\mathbf{W}\|_2^2$ where $\mathbf{U}^T$ is the $\mathbf{D}^{-1}$ Cholesky decomposition ($\mathbf{D}^{-1} = \mathbf{U}\mathbf{U}^T$). Therefore, by fixing $\mathbf{D}$, we obtain now a smooth optimization problem (8) which can be optimized directly.

$$\arg\min_{\mathbf{W},b} \sum_i L(\mathbf{Y}_i, \mathbf{X}_i\mathbf{W} + b) + \lambda tr(\mathbf{W}^T\mathbf{D}^{-1}\mathbf{W}) \quad (8)$$

To optimize (8), we adopt a direct gradient descent. Such approaches applied of the primal SVM formulation has demonstrated good performance in large scale learning setting [27]. A Quasi-Newtown LBFGS algorithm is used in this work. Compared to a classic SVM, we only need to change the definition of the derivative $\frac{\partial E}{\partial \mathbf{W}}$ to include the sparsity constraints.

$$\frac{\partial E}{\partial \mathbf{W}} = 2\sum_i (\mathbf{X}_i\mathbf{W} + b - \mathbf{Y}_i)\mathbf{X}_i + 2\lambda\mathbf{D}^{-1}\mathbf{W} \quad (9)$$

Here, $\mathbf{D}$ is an unknown variable which is dependent on $\mathbf{W}$ that also needs to be determined. We therefore use a concave-convex procedure to optimize jointly $\mathbf{D}$ and $\mathbf{W}$ in algorithm 1. Proof of algorithm 1 convergence is provided as a supplementary material due to the limitation of space.

5. Experiments

In this section we evaluate the performance of the content based pooling and WSVM model. Our approach is evaluated on three standard human action datasets: KTH, UCF50, and HMDB. Average accuracies is reported for all three datasets.

KTH [20] is composed by 6 classes of 25 human actions. The videos are subject to different zoom rates and have mostly non-cluttered static backgrounds. For the evaluation, we used the training/testing division of Schuldt [20]. UCF50 [18] is composed by 6681 video sequences distributed in 50 different human actions. Videos composing the dataset are subject to large camera motion, viewpoint change and cluttered backgrounds. In the literature two main experiment settings are used: 5 or 25 folds leave-one-out group-wise crossvalidation. To have a complete comparison with previous works, we evaluate our approach using both settings. HMDB [10] is composed by 6849 video clips divided into 51 action categories. They are collected from various sources, mostly from movies, and from public website. The different actions have large appearance variation. We adopt the default training and testing splits [10].

5.1. Experimental Setting

Dense trajectories have recently shows state-of-the-art performance for human action recognition [25]. They are therefore used as the building block of our video signature.

	KTH	UCF50 5 folds	UCF50 25 folds	HMDB
BoF	93.7	86.7	85.3	37.1
Co	94.0	88.0	87.3	40.8
Li	93.8	90.2	89.6	40.5
Mo	94.2	90.8	89.7	41.5
Spa	94.0	91.2	89.3	45.1
Mo + Li	94.2	91.7	90.6	45.9
Mo + Li + Co	94.4	92.5	91.3	48.5
Mo + Li + Co + Spa	**94.6**	**94.1**	**92.8**	**51.8**

Table 1: Average accuracies of BoF, Structural-BoFs, Spatial-BoF and their combinations. Mo, Li, Co and Spa correspond respectively to Motion, Light, Cornerness and Spatial.

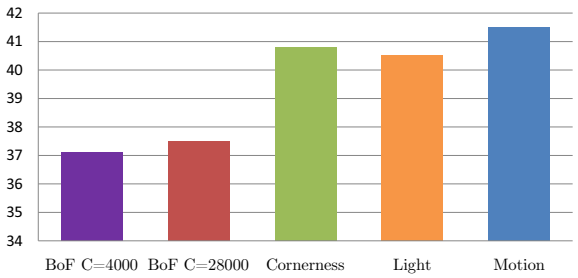


Figure 6: Impact of the dimension. C indicates the BoF codebook size. Average accuracy is reported.

To characterize a trajectory feature, motion vectors HoG, HoF and MbH descriptors are computed (see [25]) and concatenated into one vector. Since a trajectory spans on several video frames, the average saliency value of its points defines the saliency value associated to the feature. To obtain codes from trajectories, we take advantage of locality constrained coding (LCC) [13] by restricting the probabilistic soft coding to the 10 nearest words. A codebook of size 4000 is used in this experiment. We segment the saliency space with 1, 2 and 3 cells segmentation grids. We also consider spatial pooling using 2x2x2 and 3x3x3 segmentation grids [11]. The distribution of trajectory features in each spatial grid cell defines a *spatial primitive*. To combine saliency and spatial pooling, we concatenate their respective *spatial* and *structural primitives* prior to the classification. When they are not specified, WSVM parameters are set as $\lambda = 1$ and $p = 1.5$. Those values have empirically demonstrated robust performances across the different datasets. We adopt a one-versus-all classification scheme.

5.2. When Do Structural Cues Help for Action Recognition?

In a first experiment, we compare our novel pooling scheme to a traditional BoF [23] using LCC coding and max pooling. We denote the representation resulting from our content-based pooling as structural BoFs.

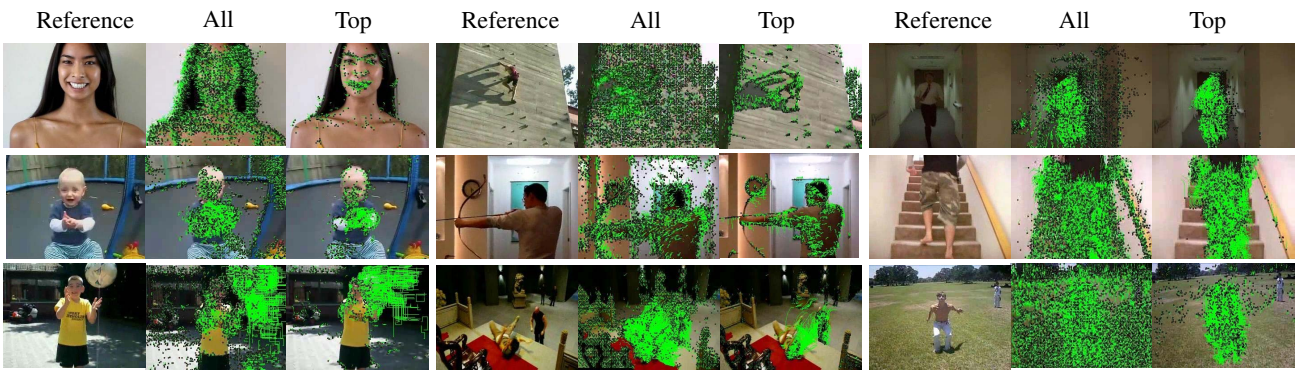


Figure 7: Illustration of prominent $\mathbf{W}_g$ groups in $\mathbf{W}$. The left column contains the reference frames. The middle column shows the extracted trajectories. The right column represents only the trajectories associated to the action most relevant *structural primitive*, i.e., the trajectories associated with the group $\mathbf{W}_g$ having the highest $\|.\|_2$ norm in $\mathbf{W}$. The most relevant *structural primitive* can be associated with cornerness, motion or light saliency depending on the action.

Results are reported Table 1. We observe that each structural BoF always outperforms traditional pooling on each dataset. Moreover as shown in Figure 6, the performance improvement is not due to the increase of the signature dimensionality. Compared to BoF with the same dimensions, structural BoFs still result in better performance. This confirms the non uniform distribution of discriminative information in the saliency spaces. By capturing the feature distribution at different saliency levels, we preserve that information in our final representation.

Motion has on average the best performance compared to the other structural BoFs. However, if we consider the accuracy per actions illustrated in Figure 10, we actually observe that the different saliencies are complementary.

For example, cornerness obtains the best performances for the actions *Smile*, *Smoke*, *Eat*. As described by Kuehne [10], those actions are characterized by close-up face views. Cornerness focuses on visually distinctive local features. In this case, it highlights features located around the noise, eye or mouth area (Figure 7). Cornerness is also useful for actions such as *Catch*, *Golf* involving objects with relatively small ellipsoidal shape.

Light gets the best performances for the actions *Climb*, *Fall Floor* or *Shooting Bow* where an upper human body is present [10]. Light saliency performs a coarse segmentation which groups together the features associated to the human body in those actions (Figure 7).

Motion achieves the best performance on actions which are characterized by a strong motion (*Chew*, *Run*, *Flic Flac*...) where the local features having high motion saliency values are likely to be part of the action of interest (Figure 7).

More generally, a structural BoF achieves significant performance improvement over a representation ignoring the space-time context when the pooling of the high saliency features only reduces the impact of the background clutter and leads to more discriminative signature.

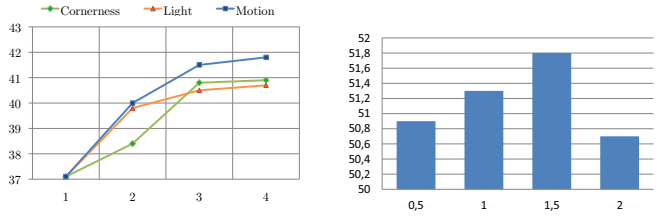
5.3. Are the Saliencies Complementary?

In this second experiment, we evaluate the combination of the different structural BoFs through the WSVM.

Table 1 reports the average accuracies of the spatial BoF and the structural BoF combination. On the HMDB dataset, an impressive performance gain of more than 16%, from 41.5 to 48.5, is achieved by the structural BoF combination (Co+Li+Mo) compared to the best individual structural BoF (Mo). This demonstrates the complementarities of saliency based representations. Furthermore, by adding spatial BoF to our video signature, another improvement of 6% is obtained. Hence, spatial and structural BoFs capture complementary information. The same trend can be observed on the UCF50 dataset. In the 25 fold setting, the combination of structural BoFs achieves an average accuracy of 91.3 compared to 89.7 for Mo. By adding spatial information, we reach 92.8.

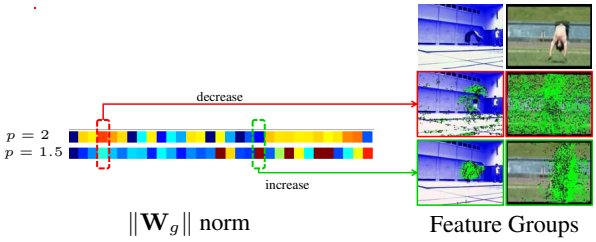
On the KTH dataset, structural BoFs as well as their combination only slightly improve over the traditional and spatial BoF. Structural BoF combination achieves a performance of 94.6 compared to 93.7 for a traditional BoF. KTH videos have almost static videos with no clutter. Most of the extracted features correspond to the foreground action, i.e. most of them are relevant to the action. It limits the need of modeling the space-time context. It should be noticed that spatial-BoF provides also a very limited improvement on this dataset, 94.0 against to 93.6.

Finally, as Table 1 shows, structural-BoF combination (Co+Li+Mo) always outperforms the spatial-BoF for each



(a) Impact of pyramid level number (b) Impact of p on structural BoFs combination

Figure 8: Parameters evaluation on HMDB. Average accuracy is reported.



$\|\mathbf{W}_g\|$ norm Feature Groups

Figure 9: Sparse feature weighting illustration.On the left, $\|\mathbf{W}_g\|_2$ are displayed, for $p = 2$ or 1.5. On the right, features corresponding to $\mathbf{W}_g$ groups are shown.

KTH		UCF50 5 fold		UCF50 25 fold		HMDB	
Laptev *et al.* [11]	91.8	Laptev *et al.* reported in [19]	47.9	Kliper *et al.* [8]	72.6	Kuehne *et al.* [10]	23.0
Wang *et al.* [25]	94.2	Sadanand *et al.* [19]	57.9	Solmaz *et al.* [24]	73.7	Sadanand *et al.* [19]	26.9
Gilbert *et al.* [4]	94.5			Reddy *et al.* [18]	76.9	Cao *et al.* [2]	27.8
Kovashka *et al.* [9]	94.5			Wang *et al.* [25]	84.5	Wang *et al.* [25]	48.3
Our approach	**94.6**	Our approach	**94.1**	Our approach	**92.8**	Our approach	**51.8**

Table 2: Comparison with state-of-the-arts. Average Accuracy is reported.

dataset showing the importance of space-time robustness.

Based on WSVM, we represent visually the trajectory features corresponding to the $\mathbf{W}_g$ having the most impact for specific actions in Figure 7.

5.4. Comparison with State-of-the-Art

Table 2 compares our approach with the state-of-the-art on each dataset. Compared to the dense trajectories BoF [25] which obtained the best performance on UCF50 25 fold and HMDB, our combination of structural and spatial BoF obtains a gains of performance of respectively 11% and 10%. On UCF50 5 fold, a strong improvement of 62% is obtained relatively to action bank [19] which had the previous best performance. It should be noticed that we also outperform action bank, from 26.9 to 51.8, on the HMDB dataset. Finally, we achieve state-of-the-art performance on KTH with an average accuracy of 94.6 compared to 94.5.

5.5. Parameters Evaluation

Figure 8a evaluates the influence of the pyramid level number on HMDB. Adding more levels increase the performance up to a certain point. To limit the dimension of our signature, we use 3 pyramidal levels in this works.

Figure 8b evaluates the impact of the sparsity parameter p on the HMDB dataset. When $p = 1.5$, WSVM outperforms a SVM ($p = 2$) from 50.7 to 51.8. While most of the performance gain comes from the saliency pooling (see Table 1), WSVM has a positive contribution of 2.1% compared to a standard SVM. For $p \leq 1$, we observe a performance decrease. In this case, $\mathbf{W}$ becomes too sparse, selecting too few *structural primitives*. It justifies the use using a $\|.\|_{2,p}$ regularizer, allowing to control the sparsity, instead of a more rigid $\|.\|_{2,1}$ norm. Figure 9 illustrates the impact of the sparsity parameter p for the HMDB "Flic Flac" action showing that p allows discriminative features to increase in importance while reducing the impact of noisy feature groups.

6. Conclusion

This paper has introduced a new space-time invariant pooling scheme that leverages the video space-time context. It identifies prominent regions in videos content through *motion*, *illumination* and *cornerness* saliencies, leading to a "video-based" segmentation scheme. We also propose a new weighted SVM that automatically learns the optimal saliency layout associated to an action. We show through an extensive experimentation that being robust to the space-time variance helps for action recognition. Our video signature combining the motion, light, cornerness, saliency and fixed spatial segmentation outperforms the state-of-the-art performances on the challenging UCF50 and HMDB datasets. In future work, we plan to investigate semantic information embedding in the pooling operation.

References

[1] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*. IEEE, 2010. 3

[2] L. Cao, Y. Mu, A. Natsev, S.-F. Chang, G. Hua, and J. R. Smith. Scene aligned pooling for complex video recognition. In *ECCV*. Springer, 2012. 2, 7

[3] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image Analysis*, 2003. 3

[4] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *CVPR*. IEEE, 2010. 7

[5] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi. Discriminative spatial pyramid. In *CVPR*. IEEE, 2011. 2, 3

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998. 3

[7] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*. IEEE, 2012. 2, 3

[8] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*. Springer, 2012. 7

[9] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*. IEEE, 2010. 7

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*. IEEE, 2011. 5, 6, 7

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*. IEEE, 2008. 1, 2, 5, 7

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*. IEEE, 2006. 1, 2

[13] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*. IEEE, 2011. 5

[14] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 2004. 2, 3

[15] F. Moosmann, D. Larlus, and F. Jurie. Learning saliency maps for object categorization. In *ECCV*. Springer, 2006. 2

[16] D. Parikh and T. Chen. Determining patch saliency using low-level context. In *ECCV*. Springer, 2008. 2

[17] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *ECCV*. Springer, 2010. 2, 3

[18] K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *MVA*, 2012. 5, 7

[19] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*. IEEE, 2012. 7

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*. IEEE, 2004. 5

[21] F. Shahbaz Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *CVPR*. IEEE, 2009. 2

[22] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*. IEEE, 2012. 2, 3

[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *CVPR*. IEEE, 2003. 1, 5

[24] B. Solmaz, S. M. Assari, and M. Shah. Classifying web videos using a global video descriptor. *MVA*, 2012. 7

[25] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 2013. 5, 7

[26] Z. Wang, B. Fan, and F. Wu. Local intensity order pattern for feature description. In *ICCV*. IEEE, 2011. 2

[27] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*. IEEE, 2009. 4, 5

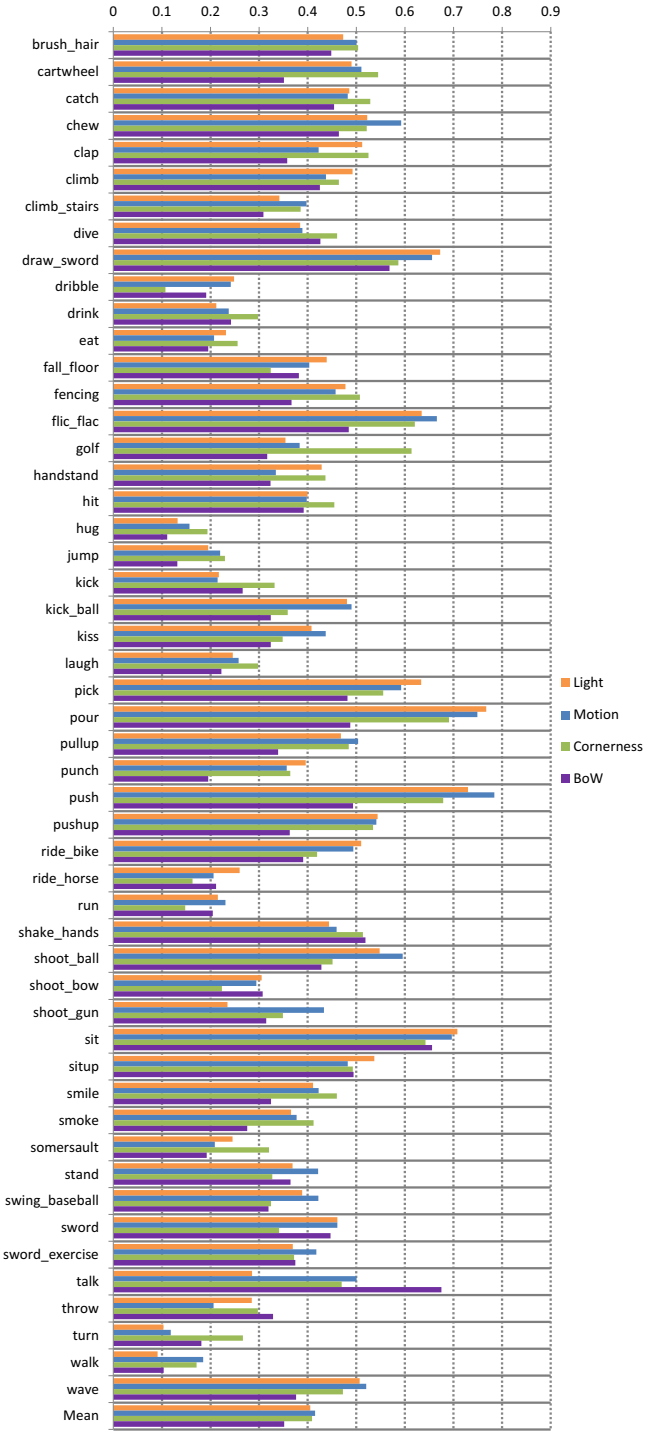


Figure 10: Per action average accuracy on HMDB.