Selection and Context for Action Recognition

Dong Han
University of Bonn
han@ins.uni-bonn.de

Liefeng Bo
TTI-Chicago
blf0218@tti-c.org

Cristian Sminchisescu
University of Bonn
sminchisescu.ins.uni-bonn.de

Abstract

Recognizing human action in non-instrumented video is a challenging task not only because of the variability produced by general scene factors like illumination, background, occlusion or intra-class variability, but also because of subtle behavioral patterns among interacting people or between people and objects in images. To improve recognition, a system may need to use not only low-level spatio-temporal video correlations but also relational descriptors between people and objects in the scene. In this paper we present contextual scene descriptors and Bayesian multiple kernel learning methods for recognizing human action in complex non-instrumented video. Our contribution is threefold: (1) we introduce bag-of-detector scene descriptors that encode presence/absence and structural relations between object parts; (2) we derive a novel Bayesian classification method based on Gaussian processes with multiple kernel covariance functions (MKGPC), in order to automatically select and weight multiple features, both low-level and high-level, out of a large collection, in a principled way, and (3) perform large scale evaluation using a variety of features on the KTH and a recently introduced, challenging, Hollywood movie dataset. On the KTH dataset, we obtain 94.1% accuracy, the best result reported to date. On the Hollywood dataset we obtain promising results in several action classes using fewer descriptors and about 9.1% improvement in a previous benchmark test.[1]

1. Introduction

We study the problem of action recognition with emphasis on human actions in complex environments, as experienced in movies or alternative sources of un-instrumented data like personal videos on the web. Recognizing human action is difficult because humans move fast, have complex structure and clothing, and their appearance is affected by scene factors like occlusion or illumination. The field of action recognition has seen lots of progress recently [19, 10, 5, 2, 13] through a synergy of new datasets, the design of low-level feature descriptors and the use of (kernel-based) machine learning methods. It has thus become possible to tackle difficult problems like the recognition of human motion in movies, *e.g*. [19]. Despite important advances, the performance of current systems in complicated settings remains relatively low. It is not clear, to date, whether the current level of performance is due to: *(a)* an intrinsic lack of discriminative power of existing video descriptors, *(b)* a lack of training data, or *(c)* a deficiency of learning, *e.g*. the inability to select optimal feature combinations. This is by no means straightforward, as explicit enumeration (or kernel selection) in the feature power set becomes infeasible for models with more than several dimensions. Each shortcoming in *(a)–(c)* suggests potentially different paths towards a solution: *(a)* the construction of complementary descriptors that account not only for low-level spatio-temporal video correlations, but also for higher-level relations among people, their parts, or the parts of other objects, *(b)* the need to acquire more data, and *(c)* better methods to learn feature/kernel combinations and any other hyperparameters of the system in a principled way. Progress along each direction influences others, and it re-

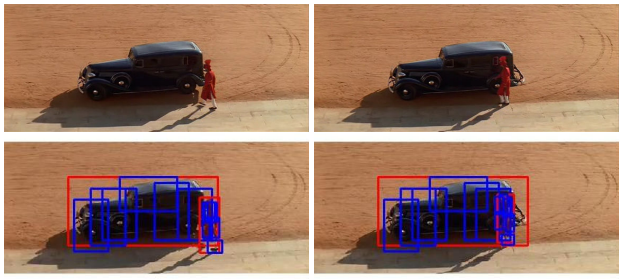

Figure 1. Illustration of multiple detected objects (car and person, in red) and their parts (in blue) on a sequence from the Hollywood dataset [19]. We aim to complement existing low-level video features with high-level contextual scene descriptors built of object detectors. By detecting both a car and a person in its proximity, the contextual recognition of an action like 'Person enters car' can be improved. A second focus of our work is the learning of large inhomogeneous descriptor ensembles, both low-level and high-level.

[1] All authors contributed equally to this research.

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

mains to be seen how much weight each component will ultimately have in a successful system.

In this paper we focus on two of the three directions just discussed: *(a)* the design of contextual scene descriptors, and *(c)* the design of more powerful classifiers. Specifically, we introduce bag-of-detector scene descriptors that encode presence/absence and structural relations between object parts. This aims to make recognition more reliable as actions like: 'person getting out of car', 'person drinking', or 'person sitting', should benefit from both the detection of a person and the object involved in an action—*e.g*. a car, a mug, a chair or a sofa—and from the knowledge of their relative spatial configurations.[2] While the use of context is not foreign to vision [21, 23, 31], its implications for action recognition are comparatively less explored [23, 34, 9, 26].

Our second contribution is the design of a novel Bayesian classification method based on Gaussian processes with multiple kernel covariance functions. This allows the principled fusion of multiple low-level and high-level kernels, out of a large collection, by means of Bayesian approximations and hyperparameter selection criteria based on marginal likelihood maximization. (To our knowledge no Gaussian process multiple kernel classification method is available at the moment either in computer vision or in machine learning.) Finally, we perform large scale evaluation on a large set of state-of-the art features and kernels on both KTH [28] and a recently introduced, complex Hollywood movie dataset [19]. We show that Gaussian process multiple kernel learning and contextual scene descriptors complement ongoing work towards understanding the importance of different feature types and the design of more accurate action recognizers.

1.1. Related work

As this research relates to feature design, kernel learning methods as well as object and action recognition, our literature review is necessarily sparse. Space-time interest points [18] use extended Harris operators to detect structures with significant local variation in spatio-temporal volumes. This sparse detector is robust against scale variations, different motion speeds and cluttered backgrounds. Combined with a bag-of-words descriptor and a discriminative classifier (*e.g*. an SVM), it offers a promising architecture for action recognition [28]. Good local descriptors need to be sufficiently discriminative to separate different classes but invariant to intra-class variations. Histogram representations possess several such properties being robust to spatial and appearance variations or illumination changes.

Dollár *et al* [4] model neighborhoods of interest points as cuboids and construct histograms based on normalized pixel intensity, brightness gradient and windowed optic flow. Laptev *et al* [19] use spatio-temporal grids extracted at multiple scales to compute histogram of oriented gradient (HoG) [3] and optic flow (HoF) within each volumetric (spatio-temporal) cell. They report state-of the art results on the KTH dataset, introduce a new, significantly more difficult movie dataset, and study the impact of different descriptors for classification performance. Scovanner *et al* [29] propose a 3D-SIFT descriptor, where gradients are computed at random locations in a cuboid and orientation histograms are built for each. This is different from HoG, in that gradients are computed in polar coordinates and histogrammed using meridians and parallels. This induces progressively smaller bins at poles. Kläser *et al* [16] refine the representation by quantizing the gradient orientation according to facets of regular polyhedrons. They improve recognition accuracy on the KTH dataset [28] using an SVM classifier with χ^2 kernel, tuned using discrete hyperparameter search. Fei-Fei Li *et al* [24, 27] propose a generative action recognition model using (unsupervised) LDA/pLSA latent topic models and features in [4] as well as correlograms. An interesting aspect of the system is that the number of action classes is determined automatically.

The work just described has made substantial progress on action recognition by focusing on low-level spatio-temporal features. In this paper we derive complementary descriptors based on quantizing responses from object detectors (bag of detectors) of the human body as well as cars or chairs. To the best of our knowledge generic bag-of-detector descriptors have not been proposed for action recognition. An interesting step in a similar direction is the recent work of [15] who segment hands and objects and fuse their correlations over time using a CRF. Applications of CRF models in conjunction with bag-of words descriptors for action recognition previously appeared in [30, 32]. Other interesting models that exploit object context for visual action analysis can be found in [23, 34, 9, 26, 22].

The second component of our contribution regards flexible classifier design. Of importance is the capacity to generalize and the flexibility to train using a large number of possibly (but not necessarily) relevant features. We will use both similarity measures *e.g*. Gaussian, histogram intersection, and feature descriptors like HoG or HoF with the goal of relevance determination for optimal design. Kernel methods are promising in this context, but in their standard form a kernel has to be pre-specified. This implicitly induces an input similarity measure and its choice makes a significant difference in the performance of the classifier. Generalizations to learning the weighting of a linear combination of kernels exist [1] and have been demonstrated to give promising results for object recognition [17] (notice that no such methods have been proposed for action recognition). A shortcoming of kernel methods is their lack of uncertainty estimates and the absence of an efficient hyper-

[2] Formulating object and action recognition in one common, jointly consistent framework remains an interesting direction for future work.

parameter learning procedure.[3] In turn, classifiers based on Gaussian processes are Bayesian methods that are more robust to overfitting, offer estimates of uncertainty (in particular, a probability distribution for the class label), and consistent hyperparameter learning by optimizing the marginal likelihood. There are a number of successful applications of Gaussian process *regression* in vision [12], but to our knowledge no full Gaussian Process *classification* method has been employed so far (in particular [12] used regression in order to successfully predict continuous values in the unit interval, for object categorization). In contrast, we follow a more integrated, consistent Gaussian process classification approach, and use the cumulative logistic function as likelihood in order to squash the latent function and obtain a meaningful distribution for the class labels. Besides, and perhaps most importantly, we develop a novel multiple kernel learning scheme in the framework of Gaussian process classification.

2. Multiple Kernel Gaussian Process Classifier (MKGPC)

For binary classification, we work with a set of inputs $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top$ and corresponding outputs $\mathbf{y} = [y_1, y_2, \cdots, y_N]^\top$, where N is the number of training samples and $y_i \in \{-1, 1\}$ the class label. Here $\mathbf{X}$ stores the inputs and $\mathbf{y}$ the class labels rowwise. The goal is to infer the class label of given inputs. We focus on the case where the covariance function of Gaussian Process is a linear combination of covariances, each with different kernel hyper-parameters, and we derive methods to learn both the weighting of different covariances and the hyperparameters of each individual kernel. We refer to this model as the *Multiple Kernel Gaussian Process Classifier (MKGPC)*.

Gaussian process classification (GPC) [25] is a discriminative model where the class membership probability $p(y|\mathbf{x})$ is Bernouilli distributed. This is achieved by squashing an unconstrained latent function $f(\mathbf{x})$ to map its values to the unit interval by a sigmoid function, *e.g.* the logistic or the cumulative Gaussian. Without loss of generality, in this paper we use the cumulative Gaussian function

$$\Phi(t) = \int_{-\infty}^{t} N(\tau|0,1)d\tau \tag{1}$$

Given the latent function, the class labels are assumed to be Bernoulli distributed and independent random variables, so the joint likelihood factors as

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f_i) = \prod_{i=1}^{N} \Phi(y_i f_i) \tag{2}$$

[3] Grid search cross validation works well in a few dimensions but does not scale to problems with tens or hundreds of dimensions that result when many kernels are combined—in this case both the weighting and the individual hyperparameters have to be estimated.

where $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)]^\top$ is a vector. A Gaussian process is a collection of random variables, any finite number of which have joint Gaussian distribution. GPC assumes that the latent function $f(\mathbf{x})$ is a zero mean Gaussian process with positive definite covariance function $k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$ which encodes correlations between input pairs, $\boldsymbol{\theta}$ are hyperparameters of the covariance. Applying Bayes' rule, we obtain the posterior over latent functions $\mathbf{f}$

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})} = \frac{\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})}{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})} \prod_{i=1}^{N} \Phi(y_i f_i) \tag{3}$$

Prediction requires two calculations: (1) the distribution of the latent variable for a given input

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \int p(f_*|\mathbf{f}, \mathbf{x}_*, \mathbf{X}, \boldsymbol{\theta}) p(\mathbf{f}|\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) d\mathbf{f} \tag{4}$$

and (2) a probability distribution for the class label using the latent function distribution in (1)

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \int p(y_*|f_*) p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) df_* \tag{5}$$

The posterior of the latent function is not Gaussian, due to the non-Gaussian nature of the likelihood (2). This makes the integrals (4) and (5) analytically intractable. A palette of integration methods including analytic approximations, *e.g.* Laplace or expectation propagation, or Monte Carlo sampling are available. For the MKGPC model, we use Laplace approximation due to its conceptual simplicity and analytic tractability.

2.1. Laplace Approximation

Laplace approximation uses a Gaussian approximation $q(\mathbf{f}|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$, derived from a second order Taylor expansion around the maximum of the posterior $p(\mathbf{f}|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$ in the integral (4), as follows

$$q(\mathbf{f}|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\widehat{\mathbf{f}}, (\mathbf{K}^{-1} + \mathbf{W})^{-1}) \tag{6}$$

where $\mathbf{W} = -\frac{\partial^2 \log p(\mathbf{y}|\mathbf{f})}{\partial \mathbf{f} \partial \mathbf{f}^\top}|_{\mathbf{f}=\widehat{\mathbf{f}}}$, and

$$\widehat{\mathbf{f}} = \mathrm{argmax}_{\mathbf{f}} \{p(\mathbf{f}|\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})\} \tag{7}$$

Notice that the $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ is independent of $\mathbf{f}$, so we only need to consider the un-normalized posterior when maximizing $\mathbf{f}$ (7). In our case, the likelihood (2) and the prior are log concave w.r.t. $\mathbf{f}$, so (7) has a global optimum. Given the factorial structure of the likelihood, the Hessian $\mathbf{W}$ is diagonal. Using (6), we can compute (4), as well as the mean and the variance of f_* analytically [25]. Based on this, we compute an approximation to the marginal likelihood

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \log \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{f} \approx \tag{8}$$

$$\approx -\frac{1}{2}\widehat{\mathbf{f}}\mathbf{K}^{-1}\widehat{\mathbf{f}} + \log p(\mathbf{y}|\widehat{\mathbf{f}}) - \frac{1}{2}\log|B|$$

where $|B| = |\mathbf{I}_N + \mathbf{W}^{\frac{1}{2}}\mathbf{K}\mathbf{W}^{\frac{1}{2}}|$ and $\mathbf{I}_N$ is the $N \times N$ identity matrix. The marginal likelihood (8) gives the probability of the data under the given model (set of parameters, hyperparameteres, *etc*.) [25] and can be used to estimate the hyperparameters $\boldsymbol{\theta}$ of the model consistently.

2.2. Multiple Kernel Covariance Function

The covariance function can be any positive definite kernel. Here we search for a linear combination of the form

$$k_m(\mathbf{x}_i, \mathbf{x}_j; \alpha, \boldsymbol{\theta}) = \sum_{t=1}^{T} \exp(\alpha^t) k(\mathbf{x}_i^t, \mathbf{x}_j^t; \gamma^t) \qquad (9)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$ are hyperparameters, $\exp(\alpha^t)$ is the weighting parameter which controls the contribution of each kernel, $\mathbf{x} = [\mathbf{x}^1, \ldots, \mathbf{x}^T]$, $\mathbf{x}^t = [x_1^t, \ldots, x_{N_t}^t]$ is the t-th group features of $\mathbf{x}$, $N = N_1 + \ldots + N_T$ and T is the number of feature groups. Notice that $k(\mathbf{x}_i^t, \mathbf{x}_j^t; \gamma^t)$ can be any positive definite kernel function, such as a Gaussian kernel, a polynomial, a (pyramid) histogram intersection kernel, *etc*. We learn multiple kernel hyperparameters α^t and γ^t by optimizing the approximated marginal likelihood (8) using an alternation scheme: we fix one block of parameters and optimize the other block, then swap the blocks, sequentially. A technical detail is the calculation of partial derivatives of (8) w.r.t. hyperparameters. The covariance matrix $\mathbf{K}$ is a function of hyperparameters, but $\widehat{\mathbf{f}}$ and $\mathbf{W}$ are also implicitly a function of $\boldsymbol{\theta}$, since as $\boldsymbol{\theta}$ changes, the optimum of the posterior $\widehat{\mathbf{f}}$ and the negative Hessian $\mathbf{W}$ change as well. The derivation is obtained using the chain rule, one component being the partial derivative of the multiple kernel function w.r.t. hyperparameters

$$\frac{\partial k_m}{\partial \alpha^t} = \exp(\alpha^t) k(\mathbf{x}_i^t, \mathbf{x}_j^t; \gamma^t) \qquad (10)$$

$$\frac{\partial k_m}{\partial \gamma_i^t} = \exp(\alpha^t) \frac{\partial k(\mathbf{x}_i^t, \mathbf{x}_j^t; \gamma)}{\partial \gamma_i^t} \qquad (11)$$

For optimization, we use the Polack-Ribiere variation of conjugate gradient with line search based on quadratic and cubic polynomial approximations. We use a slope ratio method to estimate the initial step size and a Wolfe-Powell stopping criteria [8] (see also fig. 2).

3. Image and Video Descriptors

In this section we describe the features used in experiments. We employ a variety of descriptors, both new, designed for higher level contextual scene and object modeling, *e.g*. bag-of-detectors, and mid/low-level, spatio-temporal gradient and optic flow histograms.

Bag of Detectors: We experiment with several ways to encode contextual scene information, based on object and part detections. For this purpose we use a state-of the art detector [7] tuned to identify people, both full and upper-body, as well as chairs and cars. Detectors for other objects relevant to contextual classification, *e.g*. telephone, sofa, can also be used. The detector is based on a star model with seven parts (the bounding box of detected objects is also provided) and each part is modeled using HoG descriptors. On top of this, we build 3 histogram descriptors, all normalized over sequences, that encode the presence of various objects in the scene as well as statistics of their spatial configurations: *ObjPres, ObjCount,* and *ObjDist*. *ObjPres* is a 4d descriptor that accumulates the presence/absence of each object type in the video. *ObjCount* is a 4d descriptor that counts the number of objects appearing in each image, for each category. *ObjDist* measures pair-wise distances between parts of the Person detector, accumulated for all people in the scene, normalized at frame level, and also at sequence level. Since there are 7 parts, the descriptor has dimension 21 (7 choose 2).

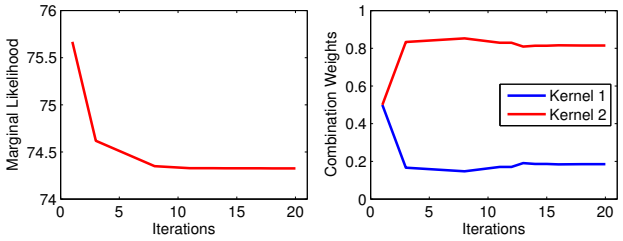

Figure 2. The progress of our conjugate gradient optimizer for a 2-kernel learning problem on the Hollywood-1 dataset. *Left*: marginal likelihood as function of iteration; *Right*: kernel weights α^1 and α^2 as function of iteration. Notice rapid convergence on this dataset.

Spatio-Temporal Gradient and Flow Features: We work with spatial-temporal interest points [18]—image locations with large spatial and temporal gradients. For scale invariance, the spatial-temporal extent of each detected event is also estimated. Instead of performing scale selection [18], interest points are extracted densely at multiple scales [19]. The motion and appearance of a space-time volume around the detected interest point at each appropriate scale is characterized in terms of features extracted on a grid of cuboids centered at the detector. HoG features (4d gradient orientations per cell) and HoF features are extracted for each image or image pair and assembled according to the desired temporal quantization level. In particular, the set of frames corresponding to the estimated scale of the interest point is split into 3, 2, and 1 (no split) sets. The HoGHoF in corresponding cells for each image or pair is then vector quantized over the assumed time window and normalized. HoG windows are typically square pixel patches of size 20, 40, or 80 pixels and the temporal frame windows are typically 10–20 frames long, depending on the spatio-temporal scale returned by the interest-point detector. We work with the following spatial-temporal descriptors (tem-

poral histograms with 1, 2, and 3 bins, *i.e.* t1, t2, t3): hog1x1 (t1, t2, t3 = 4, 8, 12d), hog3x1 (12, 24, 36d), hog2x2 (16, 32, 48d), hog2x1 (8, 16, 24d), hog4x1 (16, 32, 48d), hog3x3 (36, 72, 108d). [4] In all cases, a bag-of-words representation is generated from a codebook obtained by clustering the descriptors in the training set using k-means (we used 4000 entries).

3D gradient space-time descriptor: Similarly to [4, 18, 29], each support region in the neighborhood of interest point detectors is divided into cells using a regular grid and HoG [3] features are computed within each cell. The cuboid cells are further divided into S^3 subblocks. For each, a corresponding mean gradient is computed using integral videos. 3D orientation quantization is performed by replacing the octagon used to estimate 2D gradients in SIFT [20] with a regular polyhedron. Histograms from a support region are concatenated into a single feature vector to describe the spatial-temporal neighborhood. The spatial-temporal support is set to 8 and 6, respectively (this corresponds to image patches of size 20, 40, or 80 pixels and temporal windows 10–20 frames long, depending on the scale returned by the interest-point detector). We divide the support region into 4x4x3 cells and use an icosahedron to quantize the 3D gradient orientation to obtain a 4x4x3x20=960-d descriptor.

4. Experiments

As described in the previous section, most features (except for the bag of detectors) are extracted at spatial and temporal frames where the interest point detector [19] fires. A variety of descriptors, both bag-of-detectors and spatio-temporal features are extracted. For multiclass problems, we learn MKL classifiers for each 1-vs-all problems and use a probabilistic voting scheme in order to predict the class output. Bag-of-detector descriptors with dimension 4, 4, 21 are extracted. Spatio-temporal features hof3x3t2, hog1x1, hog3x1, hog2x2, hog2x1, hog4x1, hog3x3 all with t1, t2, t3, as well as 3D gradient and bag of detectors and features from [4], are extracted as described in §3.

KTH Dataset [28]: We compare our performance on this benchmark with the state-of-the-art.[5] The KTH dataset [28] contains six types of human actions: walking, jogging, running, boxing, handwaving, and handclapping. The actions are performed by 25 people, each several times in different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. The background is homogeneous in most sequences. There are 2391 sequences in total. We follow the experimental setup of [28] and split the dataset into training/validation set (8+8 people) and test set (9 people).

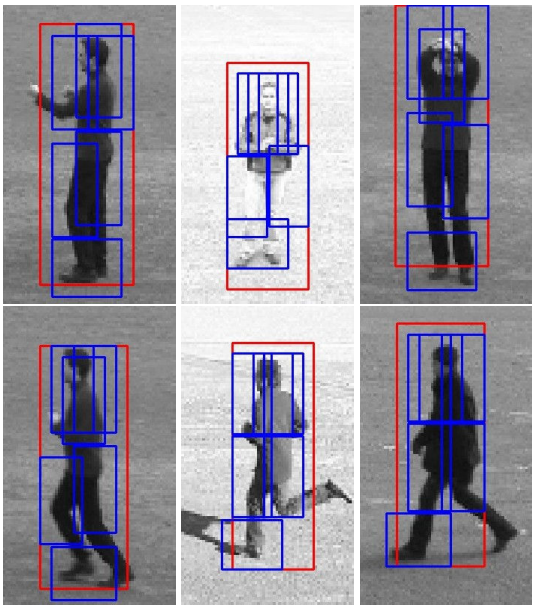

Figure 3. Illustration of the type of backgrounds, people and motions present in the KTH dataset. We also show responses of a person detector (object and parts in red and blue, respectively) on the KTH (descriptors based on these featurs were not selected by our kernel learning method, see table 1–in contrast these descriptors were found useful in the Hollywood movie dataset, see fig. 5. *Top:* boxing, handclapping, handwaving; *Bottom:* jogging, running, walking.

A comparison of our methods with the state of the art algorithm is given in table 2. The relative weights of the two selected kernels (HoGHoF and 3D Gradient) are shown in table 1. The confusion matrix of the classifier is also given in fig. 4. In this experiment hoghof3x3t2, 3D gradient features and bag of detectors are used as pool of features/kernels. Notice that we outperform the state of the art result in [19]. Our classifier tends to be superior in all motions except Running, where it achieves performance slightly inferior to [19]'s 80% accuracy (both classifiers seem to mainly confuse running and jogging, which is not entirely surprising given their visual similarity).

Hollywood-1 Movie Dataset[19]: This film dataset contains eight different action classes: answering the phone, getting out of the car, hand shaking, hugging, kissing, sitting down, sitting up, and standing up. These actions are semi-automatically collected from 32 Hollywood movies, where the training set contains video sequences from 12 movies and the testing set are sequences from other 20 different movies. In our experimental setup, we used the

[4] We use the publicly available implementation of [19] for *interest point detection*. Note that the only descriptor of [19] that is publicly available at the moment is hoghof3x3t2 (the spatio-temporal grid is hard-coded in the software). All the other spatio-temporal descriptors have been implemented by us.

[5] Note that we cannot compare against [14, 11] because their experimental setup was different (either trained on more data or splitted the problem into simpler tasks). Similarly, we can't compare against [6] as that evaluation was additionally supported by segmentation masks.

Action Category	HoGHoF	3D gradient
Boxing	0.53	0.47
Clapping	0.41	0.59
Waving	0.51	0.49
Jogging	0.43	0.57
Running	0.54	0.46
Walking	0.52	0.48

Table 1. Weighting of combining different features using MKGPC for the KTH dataset. We only show features with non-negligible weights. All the other features are weighted less than 10^{-5} during the multiple kernel learning and therefore turned-off. We suspect that the downgrade of bag-of-detector features is due to the inaccurate localization of limbs by our current detector [7]. This is a reliable, state-of-the-art human detector, but its parts do not necessarily correspond to human body parts.

Methods	Recognition accuracy
Schuldt *et al* [28]	71.7%
Niebles *et al* [24]	83.3%
Wong *et al* [33]	86.7%
Savarese *et al* [27]	86.8%
Kläser *et al* [16]	91.4%
Laptev *et al* [19]	91.8%
Ours (MKGPC)	94.1%

Table 2. Comparison of our method with the different state-of-the-art results on the KTH dataset.

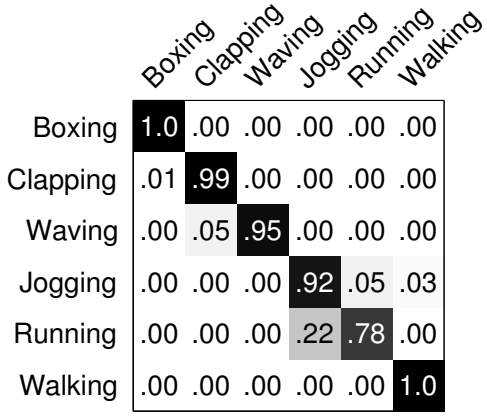

Figure 4. Confusion matrix for the KTH action dataset, obtained for a model where hoghof3x3t2 and 3D gradient features were automatically selected out of a larger set (see text).

clean training set labels (231 sequences) and clean testing set labels (217 sequences), which were manually labeled as ground truth. The dataset is huge, so to extract spatio-temporal HoGHoF we subsample the extracted descriptors to select a subset of 100,000 for clustering in order to obtain the visual vocabulary. The codebook size is 4000. Bag of words detectors are computed by only running the detectors in frames where at least one spatio-temporal interest point is detected. Samples from the dataset are shown in fig. 5 and results obtained using various descriptors are reported in table 3. Notice that we work with a significantly less extensive set of features than [19] and we obtain better performance in 5 out of 8 classes, with significant improvement for the action 'SitUp'.

Table 3 compares the average precision of our experiments with the results that were shown in [19]. The first and second columns are per-class average precision by using standard bag-of-features with HoG or HoF respectively, as reported in [19]. The third column reports the classification result obtained with the best single feature. We also show the accuracy of a random classifier (chance).

Table 4 compares again the average precision of our experiments with the results that were shown in [19]. Notice that those results were obtained by using the test set to drive a greedy selection process [19]. We have implemented one possible approximation and find the combination of kernels based on a greedy approach. Starting with an empty set, we initiate a forward selection scheme with kernel additions and removals, each evaluated, using the test error, on possible regularization parameters that consist of a kernel parameter γ and a regularization parameter C: $\{2^{-4}, 2^{-3}, \ldots, 2^{9}, 2^{10}\} \times \{2^{2}, 2^{1}, \ldots, 2^{13}, 2^{14}\}$ until a maximum is reached. This range turned to be sufficient for our problem. Notice that this is not a standard experiment, as also the authors pointed out to us [19]. It was just intended (by both [19] and us) in order to study the intrinsic power of different features. Under this experimental setting, we gain about 9.1% over reports in [19].

Hollywood-2 Movie Dataset [22]: This dataset is a recently enlarged version of Hollywood-1 with four new added actions: DriveCar, Eat, FightPerson and Run. Twelve action classes are extracted from 69 Hollywood movies, where the total length of action samples is about 600k frames. Training and test sets consist of 823 and 884 sequences, respectively, with no sharing of samples from the same movie. For preliminary experiments (as [22] is released and published at about the same time as this work), we extracted four types of features: HoGHoF, ObjPres, ObjCount and ObjDist as wel as spatio-temporal features. Our results are compared to [22]'s in table 5, where due to space limitations, for each action class, the feature with the highest weight in the kernel combination is shown.

5. Conclusions

We have presented descriptors and learning methods for action recognition. We contribute in three areas: (1) the design of descriptors based on bags of detectors, in order to encode contextual relations between people and objects in the scene; (2) a Gaussian process classifier with multiple kernel covariance function (MKGPC) that allows the principled learning of all parameters and hyperparameters

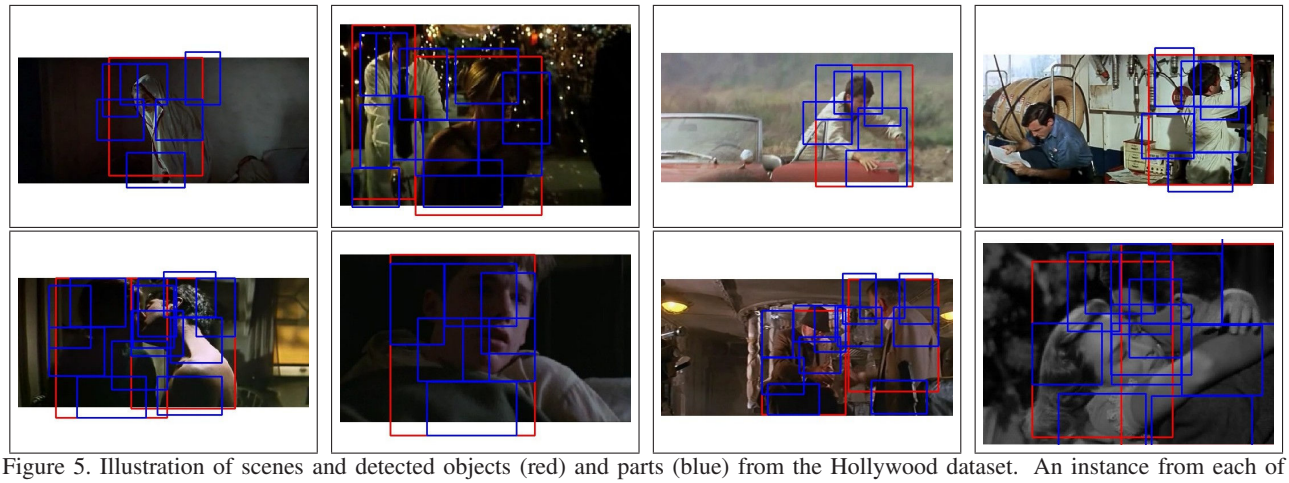

Figure 5. Illustration of scenes and detected objects (red) and parts (blue) from the Hollywood dataset. An instance from each of the action classes is shown. Top: StandUp, SitDown, GetOutCar, AnswerPhone; Bottom: Kiss, SitUp, HandShake, HugPerson. Notice that the detector is not always reliable, *e.g.* only one person is detected in the top-right image ('Answer phone' action). However, since our descriptors are based on normalized temporal histograms, it is often enough if sufficiently good responses are obtained over a subset of the frames.

Action category	HoG [19]	HoF [19]	Our best channel	Chance
AnswerPhone	13.4%	24.6%	24.1% (ObjDist)	10.6%
GetOutCar	21.9%	14.9%	24.9% (ObjPres)	6.0%
HandShake	18.6%	12.1%	26.8% (ObjDist)	8.8%
HugPerson	29.1%	17.4%	20.2% (hog2x1t2)	10.1%
Kiss	52.0%	36.5%	47.3% (hog2x2t3)	23.5%
SitDown	29.1%	20.7%	37.2% (hog2x2t1)	13.8%
SitUp	6.5%	5.7%	25.7% (hog4x1t3)	4.6%
StandUp	45.4%	40.0%	50.8% (HoGHoF)	22.6%

Table 3. Comparison of per-class average precision on different features (Hollywood-1). The best channel means the feature giving the best performance. Our results were obtained using standard model learning/fitting procedures to learn the hyperparameters that only operated with the training set (no information from the test set was used, in any form).

Action category	Baseline [19]	Ours
Answerphone	32.1% (hof o2x2t1, hof3x1t3)	43.4% (ObjCount, hog4x1t3, ObjPres, ObjDist)
GetOutCar	41.5% (hof o2x2t1, hog3x1t1)	46.8% (ObjPres, HoGHoF, hog4x1t1)
HandShake	32.3% (hog3x1t1, hog o2x2t3)	44.1% (ObjDist, ObjCount)
HugPerson	40.6% (hog1t2, hog o2x2t2, hog3x1t2)	46.9% (HoGHoF)
Kiss	53.3% (hog1t1, hof1t1, hof o2x2t1)	57.3% (HoGHoF, hog1x1t1, hog2x2t3)
SitDown	38.6% (hog1t2, hog1t3)	46.2% (HoGHoF, hog3x3t1, hog1x1t1)
SitUp	18.2%(hog o2x2t1, hog o2x2t2, hog3x1t2)	38.4% (hog4x1t3, hog3x1t1)
StandUp	50.5%(hog1t1, hof1t2)	57.1% (HoGHoF, ObjDist)
Mean	**38.4%**	**47.5%**

Table 4. Comparison of feature separation power (Hollywood-1). We implement a similar greedy selection as generally described in [19] on our feature set (see text). Notice that this is not a standard experiment because the test error is used to choose the combination of kernels. It only shows the potential of various features.

of the model using Bayesian marginal likelihood maximization procedures, and (3) a large scale study of a number of state-of-the art descriptors on the KTH and the Hollywood movie datasets. We show that the combination of descriptors achieves results superior to the state-of-art methods in two challenging datasets. A conclusion of our study is that combining contextual object detectors and multiple kernel selection techniques can be promising for the design of a reliable action recognition systems.

Future work: We work on scaling MKGPC to large multiclass problems and on the design of contextual scene descriptors that integrate robust, probabilistic object tracking, as well as an extended set of object detectors. Good margins for improving the performance of the current systems and for enhancing the diversity of existing datasets remain.

Acknowledgements: This research was supported, in

Action Category	Best channel [22]	Our best channel
AnswerPhone	10.7% (SIFTHoGHoF)	15.57%(Hog3x3t1)
DriveCar	75.0% (SIFTHoGHoF)	87.01% (HoGHoF)
Eat	28.6% (SIFTHoGHoF)	50.93% (hog3x3t2)
FightPerson	67.5% (HoGHoF)	73.08% (HoGHoF)
GetOutCar	19.1% (SIFTHoGHoF)	27.19% (ObjCount)
HandShake	14.1% (SIFTHoGHoF)	17.17% (HoGHoF)
HugPerson	13.8% (SIFTHoGHoF)	27.22% (HoGHoF)
Kiss	55.6% (SIFTHoGHoF)	42.91% (HoGHoF)
Run	56.5% (SIFTHoGHoF)	66.94% (Hog3x3t3)
SitDown	31.6% (HoGHoF)	41.61% (HoGHoF)
SitUp	14.2% (SIFT)	7.19% (Hog1x1t3)
StandUp	35.0% (HoGHoF)	48.61% (HoGHoF)
Mean	**35.1%**	**42.12%**

Table 5. Comparison of per-class average precision on different features (Hollywood-2), with the feature having the highest weight in the kernel combination shown. The results are obtained using standard model fitting procedures in order to learn the hyperparameters. We only operate on the training set; no information from the test set was used, in any form.

part, by awards from the European Commission (MCEXT-025481) and NSF (IIS-0535140).

References

[1] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, New York, NY, USA, 2004.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *ICCV*, 2005.

[3] N. Dalal, B. Triggs, and C. Schmid. Human Detection Using Oriented Histograms of Flow and Appearance. In *ECCV*, May 2006.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *VS-PETS*, 2005.

[5] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing actions at a distance. In *ICCV*, 2003.

[6] A. Fathi and G. Mori. Action Recognition by Learning Mid-level Motion Features. In *CVPR*, 2008.

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Mode. In *CVPR*, 2008.

[8] R. Fletcher. Practical Methods of Optimization. In *John Wiley*, 1987.

[9] A. Gupta and L. Davis. Objects in action: An approach for combining understanding and object perception. In *CVPR*, 2007.

[10] N. Ikizler and D. Forsyth. Searching video for complex activities with finite state models. In *CVPR*, October 2007.

[11] H. Jhuang, T. Serre, L.Wolf, and T. Poggio. A Biologically Inspired System for Action Recognition. In *ICCV*, 2007.

[12] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007.

[13] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Crowded Videos. In *ICCV*, 2007.

[14] T. K. Kim, S. F. Wong, and R. Cipolla. Tensor Canonical Correlation Analysis for Action Classification. In *CVPR*, June 2007.

[15] H. Kjellström, J. Romero, D. Martinez, and D. Kragic. Simultaneous visual recognition of manipulation actions and manipulated objects. In *ECCV*, 2008.

[16] A. Kläser, M. Marszałek, and C. Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In *BMVC*, 2008.

[17] A. Kumar and C. Sminchisescu. Support kernel machines for object recognition. In *ICCV*, 2007.

[18] I. Laptev and T. Lindeberg. Space-Time Interest Points. In *ICCV*, 2003.

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

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

[21] R. Mann, A. Jepson, and J. Siskind. The Computational Perception of Scene Dynamics. In *CVIU*, volume 65(2), pages 113–128, 1997.

[22] M. Marszałek, I. Laptev, and C. Schmid. Actions in Context. In *CVPR*, June 2009.

[23] D. Moore, I. Essa, and M. Hayes. Exploting human actions and object context for recognition tasks. In *ICCV*, 1999.

[24] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *IJCV*, 79(3):299–318, 2008.

[25] C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*. the MIT Press, 2006.

[26] M. Ryoo and J. Aggarwal. Hierarchical Recognition of Human Activities interacting with Objects. In *Semantic Learning Workshop at CVPR*, 2007.

[27] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei. Spatial-Temporal Correlations for Unsupervised Action Classification. In *WMVC*, 2008.

[28] C. Schuldt, I. Laptev, and B. Caputo. Recognizing Human Actions: A Local SVM Approach. In *ICPR*, 2004.

[29] P. Scovanner, S. Ali, and M. Shah. A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition. In *ACM Multimedia*, 2007.

[30] C. Sminchisescu, , A. Kanaujia, and D. Metaxas. Conditional Models for Contextual Human Motion Recognition. *CVIU*, 104(2-3):210–220, 2006.

[31] A. Torralba. Contextual Priming for Object Detection. *IJCV*, 53(2), 2003.

[32] S. Wang, A. Quattoni, L. P. Morency, D. Demirdjian, and T. Darrell. Hidden Conditional Random Fields for Gesture Recognition. In *CVPR*, 2006.

[33] S. F. Wong and R. Cipolla. Extracting Spatiotemporal Interest Points using Global Information. In *ICCV*, October 2007.

[34] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A Scalable Approach to Activity Recognition Based on Object Use. In *ICCV*, 2007.