Tracklet Descriptors for Action Modeling and Video Analysis

Michalis Raptis and Stefano Soatto

University of California, Los Angeles
{mraptis, soatto}@cs.ucla.edu

Abstract. We present spatio-temporal feature descriptors that can be inferred from video and used as building blocks in action recognition systems. They capture the evolution of "elementary action elements" under a set of assumptions on the image-formation model and are designed to be insensitive to nuisance variability (absolute position, contrast), while retaining discriminative statistics due to the fine-scale motion and the local shape in compact regions of the image. Despite their simplicity, these descriptors, used in conjunction with basic classifiers, attain state of the art performance in the recognition of actions in benchmark datasets.

1 Introduction

The analysis of "activities" (or "events" or "actions") in video is important and yet elusive as there is no obvious taxonomy and their measurable correlates are subject to significant variability. While many activities can be classified based on still images [33], the temporal evolution is important to tease apart more subtle differences [12]; it is obvious that a viable approach has to successfully combine both spatial and temporal statistics. We use the words "activities" or "actions" in quotes, because we do not have a precise (operational) definition for them. However, we postulate that such complex phenomena can be understood as the composition of relatively simple *spatio-temporal statistics*, which we will attempt to characterize in Sect. 2.

In this paper we define elementary spatio-temporal statistics under a set of modeling assumptions about the image formation process (Sect. 2), propose a model to infer them (Sect. 2.2), and evaluate the resulting descriptors on classification tasks using benchmark datasets (Sect. 4).

We *focus on low-level representation,* to devise statistics of the spatio-temporal signal that are insensitive to nuisance factors and yet sufficiently discriminative, that can be used as *building blocks* for more sophisticated models that exploit top-down structure and priors. Thus we purposefully operate with impoverished models that emphasize the low level, keeping top-down processing, shape and motion priors, and learning machinery to a minimum. Even with this impoverished representation, we show that we can achieve competitive performance in end-to-end classification tasks on benchmark datasets. More importantly, however, we believe that our features can be profitably used by more sophisticated models that do exploit top-down information in the form of global temporal statistics or spatial context.

1.1 Related work

We propose spatio-temporal feature descriptors that capture the local structure of the image around trajectories tracked over time. We actively *restrict* our attention to a subset of the spatial image domain and encode its "local photometry". Our approach differs from "holistic" ones [3, 8, 43, 20, 42] that use the entire video volume to extract global statistics, and compare them with standard norms, block correlation [43], or dynamic time warping [20]. Unlike these approaches, we explicitly model "simple" nuisance variability (position, contrast etc.), detect a corresponding frame with a co-variant detector, and "undo" it in the descriptor, which is therefore by construction invariant to such nuisances. The residual "complex" nuisances (local deformation, deviation from Lambertian reflection, complex illumination changes) are instead averaged out in the descriptor. Such averaging is performed relative to the structure of the nuisances, learned during the training phase, and plays a similar role to spatial binning (a form of "unstructured" averaging) in [23]. In this sense, our approach relates to part-based representations for action recognition, including [34, 7, 21, 29, 40].

Different local descriptors have been proposed to capture shape [34, 7] or joint motion and shape [18, 17, 4] by aggregating features within video cubes centered at spatio-temporal interest points into a static descriptor. In contrast, we retain in our *tracklet* descriptor the entire feature time series from birth to death of each tracked region. Other recent works [38, 27, 25, 13] also use a collection of trajectories to increase the discriminative power of local spatio-temporal volumes, but utilize different representations: [38] uses the stationary statistics of the Markov chain of instantaneous velocities to describe the evolution of the trajectories, which suffers from small-sample effects, while we explicitly maintain the entire time series and employ dynamic time warping to compare our variable-length descriptors. Messing *et al.* [27] use velocities as observations in a sequential graphical model.

We illustrate the general architecture of our descriptors using off-the-shelf detectors and local motion estimators and perform averaging or aggregation using the computational architecture of [23]. While more sophisticated instantiations are possible, already these simple choices attain state-of-the-art performance in the Activities of Daily Living (ADL)[27], the KTH [34] and the Hollywood Human Action (HOHA) [17] datasets. The implementation of the proposed descriptor is available at: `http://vision.ucla.edu/~raptis/tracklets`.

2 Spatio-temporal Tracklet Descriptors

We now describe the modeling assumptions under which we operate, and the procedure to infer the resulting representation (Sect. 2.2). While one would want to assemble these elementary actions (dictionary elements) into a model that captures the joint spatio-temporal statistics at a more global spatial scale ("context"), in Sect. 4 we show that even a naive use of the dictionary labels as a "spatial bag" yields competitive performance in end-to-end tasks.

2.1 Model and assumptions

We assume that each "object" is defined *at rest* as a compact region of space, only part of which may be visible due to occlusions, and projected onto a subset D of the image plane, yielding a function $\rho : D \subset \mathbb{R}^2 \to \mathbb{R}^+;\ x \mapsto \rho(x)$ where $D \subset \Omega$ is the *base image region*. There is no requirement that an entire object be captured by one base region. Instead, we can expect objects to be over-segmented in multiple base regions, with their spatio-temporal relations characterizing the object.[1] Base regions move under the action of a finite-dimensional group $g(t) \in \mathbb{G}$, which we assume without loss of generality to be the group of rigid motions $\mathbb{G} = SE(2)$, with the residual motion, that depends on the shape of the scene and viewpoint, captured by a general diffeomorphism $w : \Omega \to \Omega; x \mapsto w(x)$. Finally, a contrast transformation is applied to the range of the image in the base region, and all other photometric factors (specularities, translucency, interreflections etc.) are lumped together as an additive component $n(x,t)$. These assumptions are summarized in the model:

$$\begin{cases} \rho(x),\ x \in D \subset \mathbb{R}^2 \quad \text{base region} \\ \rho \circ g(t) \doteq \rho(g(t)x),\ g(t) \in SE(2) \quad \text{global motion} \\ \rho \circ w(x,t) \circ g \doteq \rho\left(w(g(t)x,t)\right)\ \ w : \mathbb{R}^2 \to \mathbb{R}^2 \quad \text{local deformation} \\ h(t) \circ \rho \circ w \circ g \doteq h(\rho(w(g(t)x,t)),t), \quad h : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \quad \text{contrast} \\ I(x,t) = (h \circ \rho \circ w \circ g)(x,t) + n(x,t) \quad \text{complex illumination, noise, etc.} \end{cases} \tag{1}$$

The above equation is valid only for those $x \in \mathbb{R}^2$ that intersect the domain of the image Ω. Elsewhere, the image is due to phenomena other than the base region, which we call *clutter*, $\beta(x,t)$. So, the actual measured image is given by

$$I(x,t) = \begin{cases} h \circ \rho \circ w \circ g(x,t) + n(x,t), \qquad \forall\ x \in g^{-1}(t)w^{-1}(D,t) \cap \Omega \\ \beta(x,t) \quad \text{elsewhere.} \end{cases} \tag{2}$$

The *components* (hidden factors) of the extended temporal observation of an object are the (multiple) base image regions $\rho^i_{|D}$, their (variable) length $\hat{T}_i = T_i - \tau_i$, global trajectory $\{g_i(t)\}_{t=\tau_i}^{T_i}$, their local deformation[2] $\{w_i(x,t);\ x \in g_i(t)D\}_{t=\tau_i}^{T_i}$, the contrast transformation $\{h_i(t)\}_{t=\tau_i}^{T_i}$, while everything else is lumped in $n_i(x,t)$. In the rest of this section we will omit the index i and focus on *inference and representation*: How can we "extract" the hidden components from a time series $\{I(x,t), x \in \Omega\}_{t=\tau}^{T}$? What components of the data-formation process matter for classification? In order to make the inference tractable, we make the following *modeling assumptions*: The effect of complex nuisances $n(x,t)$ is small relative to other factors, so we **(a)** seek explanations of the data that minimize their effects (e.g. a suitable norm of $n(x,t)$). The contrast

[1] Although it is precisely these contextual spatial relations that we ignore in Sect. 4, to test the representational power of the descriptor alone.

[2] Here $g_iD = \{g_ix \mid x \in D\}$.

transformation h "contains no information" (i.e., we wish the outcome of the task to be independent of contrast), so we **(b)** seek to eliminate it from the representation. The global motion $g(t)$ *may or may not* contain information, depending on the task, so we **(c)** seek to infer it from the data for later use, or to **(d)** provide a local reference where to compute the deformation field $w(x,t)$. The base region ρ and the local deformation w contain all the photometric, geometric and dynamic information, respectively, embedded in the data. Therefore, the *inference problem* can be stated as:

$$\{\hat{\rho}, \hat{w}, \hat{g}\}_{t=\tau}^{T} = \arg\min_{\rho,w,D,h,g} \int_{\tau}^{T} \|n(x,t)\|_D dt \tag{3}$$

subject to (2), where $\|n(x,t)\|_D = \int_D |n(x,t)|^2 dx$, with the addition of an area regularizer to avoid the trivial solution $D = \emptyset$. This formalizes (a). To eliminate h, (b) we simply encode the estimate of the base image region $\hat{\rho}_I(x) \doteq I \circ \hat{w}^{-1} \circ \hat{g}^{-1}$ using a complete contrast-invariant, such as the geometry of the level lines (or its dual, the gradient orientation), or a local contrast normalization, e.g.

$$\phi(\hat{\rho}(x)) \doteq \frac{\nabla \hat{\rho}_I(x)}{\|\nabla \hat{\rho}_I(x)\|_\epsilon} \quad \text{or} \quad \phi(\hat{\rho}(x)) \doteq \frac{I - \int_D I dx}{\|\text{std}(I_{|_D})\|_\epsilon} \tag{4}$$

where[3] $\|I\|_\epsilon = \min\{\|I\|, \epsilon\}$. We are then left with estimating (c) the global motion g, and (d) the local deformation w. Rewriting eq. (3) we have a sequence of equivalent optimizations in fewer and fewer unknowns:

$$\begin{aligned} &\arg\min_{h,\rho,w,g} \int_{\tau}^{T} \int_D |I(x,t) - h \circ \rho \circ w \circ g| dx dt = \quad \text{(thm. 7.4, p. 269 of [31])} \\ &= \arg\min_{\rho,w,g} \int_{\tau}^{T} \int_D |\phi(I(x,t)) - \phi(\rho \circ w \circ g)| dx dt = \; \text{(thm. 1, p. 4 of [37])} \\ &= \arg\min_{w,g} \int_{\tau}^{T} \int_D |\phi(I(x,t)) - \phi(I(x,t+1) \circ w \circ g)| dx dt \doteq \{\hat{g}(t), \hat{w}(x,t)\} \end{aligned} \tag{5}$$

This problem can be solved using variational optimization techniques [37]; a more efficient, albeit suboptimal, solution can be arrived at by first assuming $w(x,t) = x$ and estimating $\hat{g}(t) = \arg\min_g \int_\tau^T \int_D |\phi(I(x,t)) - \phi(I(x,t+1) \circ g(t))| dx dt$ with any tracking algorithm [24, 35, 27]. Then, given $\{\hat{g}(t)\}_{t=\tau}^T$, estimate $\hat{w}(x,t) = \arg\min_w \int_\tau^T \int_D |\phi(I(x,t)) - \phi(I(x,t+1) \circ w \circ \hat{g}(t))| dx dt$ with any optical flow algorithm. Note that $\hat{w}$ depends on $\hat{g}$, and there is no guarantee that substituting $\hat{w}, \hat{g}$ in (5) minimizes the cost. However, this approach is sufficient for our purposes, otherwise one can revert to an infinite-dimensional optimization of (5).

[3] Since the gradient direction will be weighted by its norm in the averaging operation to compute the descriptor (Sect. 2.2), the value of ϵ does not matter in practice. As an alternative, when color images are available, one can use spectral ratios or local normalization to eliminate contrast transformations.

2.2 Simplest instantiation and inference of the representation

Following the derivation above, given a video sequence $\{I(x,t),\ x \in \Omega\}_{t=1}^{T}$, we first select candidate regions via any feature detector [23, 10, 1], and track them over time using a contrast-compensated translational tracker to obtain a number of trajectories $\{\hat{g}_i(t)\}_{t=\tau_i}^{T_i}$ of varying length $\hat{T}_i$, addressing (c). Many trackers also provide a rotational and scale reference; the latter can be used to select the base regions $D_i \subset \mathbb{R}^2$. The former can be used to fix local orientation, although we select the vertical image coordinate as reference. In the resulting local frame $\{D_i, \hat{g}_i(t)\}$ we then estimate the local motion $\{\hat{w}_i(x,t)\}_{t=\tau_i}^{T_i}$ using any of a number of local optical flow algorithms, the simplest being [24]. This addresses (d) and completes the (co-variant) frame selection process. Therefore, we design an invariant descriptor by representing the image in the selected frame, $\{D_i, \hat{g}_i(t)\}$ via the contrast invariant $\{\phi(I \circ \hat{g}_i)\}$, and concatenate that with the motion field $\{\hat{w}_i(x,t) \circ \hat{g}_i(t)\}$ in the base region D_i.

If we had priors on the intra-class variability $dP(g,w)$, we would marginalize the resulting descriptor; in their absence, it is common to assume that the object or category of interest is described by an "uncertainty ball" around a reference descriptor, that is therefore "blurred" in some sense, ideally by averaging with respect to the prior, but more often by coarse spatial binning. In the latter case, the descriptor for $\{\phi(I \circ \hat{g}_i)\}$ corresponds to a histogram of gradient orientations (HoG) [23, 6], and the descriptor for $\{\hat{w}_i(x,t) \circ \hat{g}_i(t)_{|D_i}\}$ corresponds to a histogram of optical flow vectors (HoF).

Although many have used HoG/HoF descriptors [18, 22, 17, 4], they aggregate them into a static signature, whereas our previous analysis and [36] suggest retaining their temporal evolution. However, rather than averaging by spatial binning (that presumes ergodicity), we prefer to use at least a crude approximation of the prior $dP(g,w)$ in the form of samples $\{g(t_j)\},\ \{w(x,t_j)\}$ inferred during the training phase. The resulting descriptor, which we call AoG (average of gradient orientation) and AoF (average of optical flow), averages over the training samples – inferred in a sliding temporal window $\{t_j\}_{j=1}^{L}$ and thought of as samples from an importance distribution:

$$AoG(t|x, g_i, D_i) = \sum_{\tau=t-\lfloor L/2 \rfloor}^{t+\lfloor L/2 \rfloor} \phi(I(x,\tau)) \circ g_i^{-1}(\tau) \qquad x \in g_i(\tau) D_i \cap \Omega \tag{6}$$

where $g_i D_i$ is defined in footnote 2. Although "oG" in AoG stands for the gradient orientation, in analogy to HoG, any other contrast-normalizing statistic ϕ can be used, as in (4). Similarly, we have

$$AoF(t|x, g_i, D_i) = \sum_{\tau=t-\lfloor L/2 \rfloor}^{t+\lfloor L/2 \rfloor} (w_i \circ g_i)(x,\tau) \qquad x \in D_i \cap \Omega \tag{7}$$

We call *Tracklet Descriptor* (TD) the concatenation of the entire time series of either HoG/HoF, or AoG-HoF, and compare the two in Sect. 4, where we show the latter to yield marginally improved performance at a significantly lower computational cost. Optionally, the TD can be augmented with some sample statistic, for instance the trajectory relative to the spatial or spatio-temporal mean.

$$\boxed{\pi_i(t|I) \doteq \{A/HoG_i(t), A/HoF_i(t)\}} \tag{8}$$

As stated in Sect. 1, we postulate *compositionality* of our representation, so it is natural to organize tracklet descriptors into a "dictionary." However, because we retain the entire time series, the process is more involved as descriptors of different length have to be compared. In Sect. 3 we describe how this can be done using dynamic time warping and clustering by affinity propagation. As an alternative to averaging, one could consider histograms aggregated over time, rather than space, with similar results, as advocated by [19].

3 Implementation

Following Sect. 2.2, we reduce the group $\mathbb{G} = \mathbb{R}^2$ to pure translations, and estimate $\{\hat{g}_i(t) \in \mathbb{G}\}_{t=\tau_i}^{T_i}$ using [35], as implemented by [2], without affine consistency check, similar to [27]. Features lost during tracking are replaced by newly selected ones. We prune tracks that are less than T_i = 5-frames long, or that move less than $\hat{g}_i(T_i)$ = 3-pixels in standard deviation. Unlike [38], we do not impose an upper bound on $\hat{T}_i$, and unlike [7, 34, 4, 25] we do not use a fixed time-scale.

3.1 Constructing tracklet descriptors

We capture the contrast-invariant statistics ϕ of the base regions D_i using the gradient orientation spatially binned (HoG) or averaged (AoG) in a sliding temporal window, e.g., $L = 5$ with fixed scale and orientation, centered at each spatial location $\hat{g}_i(t)$ along the trajectory. The size of $D_i(t)$ could be adapted using the scale component estimated on-line by the tracker. Although we estimate rotation of the base regions D_i we discard it, and use the vertical component of the image plane as a reference. In yet a simpler instantiation, one can consider the base regions D_i fixed to, say, 18×18 or 32×32 pixels. We estimate the local deformation $\hat{w}_i(x, t)$ using [24] and aggregate it either in a spatial histogram (HoF) or in an average (AoF) within each region D_i. While HoG/HoF result in a fixed 128-dimensional vector each, AoG/AoF have variable size depending on $|D_i|$; therefore they are quantized into a comparable number of components (225 in the experiments, corresponding to 15×15 patches). The two vectors are concatenated[4] and stacked sequentially over time into a matrix.

3.2 Tracklet dictionary

For each base image region D_i, a tracklet descriptor represents a multi-dimensional time series, $\pi_i : [\tau_i, T_i] \to \mathbb{R}^N$. To define a distance between two descriptors we

[4] Although one could introduce weights between the spatial and temporal component, and optimize the weight to a particular dataset, we do not do so in Sect. 4.

must discount initial time, speed of execution, and duration of an action. Therefore, we adopt the dynamic time warping (DTW) distance [32]:

$$d(\pi_i, \pi_j) \doteq \inf_{\alpha,\beta \in \mathcal{H}} \frac{1}{M} \sum_{t=1}^{M} \|\pi_i(\alpha(t)) - \pi_j(\beta(t))\|_1 \tag{9}$$

where $\alpha, \beta \in \mathcal{H}$ are continuous monotonic transformations [39, 20] of the temporal domain. For HoG, HoF and AoG we use the ℓ_1 distance. Optical flow vectors, however, are not sparse, so ℓ_2 should be used instead, allowing small discrepancies. Therefore, AoG and AoF cannot be simply concatenated, but instead separate dictionaries, and combinations of separate kernels, have to be learned. The different structures of AoG and HoF also do not lead to a "meaningful" compact descriptor. To make comparison as fair as possible, in Sect. 4 we test AoG vs. HoG in isolation (Table 3). For a track of 100 frames, HoG takes 13 seconds to be computed (in non-optimized C code), whereas AoG takes 0.6 seconds (in Matlab).

Because of the variable length, many commonly used clustering algorithms (e.g., k-means) are inapplicable to clustering time series. Agglomerative clustering [15] and k-medoids have been used to select cluster centers for time series. We compute pairwise distances among tracklet descriptors, and set the distance to infinity for pairs with length ratio not between 0.5 and 2, since DTW does not provide a meaningful warping path for those cases [30]. We use affinity propagation [9] to cluster and select dictionary elements. This method is efficient due to the sparsity of the initial distance matrix and effective to define discriminative exemplars without the need of multiple random initializations that algorithms like k-centers require. In our experiments the size of the dictionaries was not pre-specified but it was automatically selected by affinity propagation.

It is not immediate to visualize our cluster centers, since our model is not strictly generative. However, Fig. 1 shows parts of the tracks colored according to their nearest neighbor in a tracklet dictionary. Fig. 2 shows a sample trajectory with samples of the quantized histogram of gradient orientations and optical flow super-imposed on the image. These histograms are concatenated to form a temporal sample of the time series $\{\pi_i(t)\}$.

3.3 A basic classification scheme

The simplest recognition method we consider is akin to a bag-of-features (BoF) [5], whereby we discard global temporal ordering, capturing only the local temporal variation of a tracklet. This admittedly naive model achieves performance already close to the state of the art. Given a codebook of TDs, we assign each trajectory in a test frame to the closest codebook element (Sect. 3.2); then each video is represented by a histogram of occurrences of dictionary elements. We use a support-vector machine with either a RBF-χ^2 kernel or an intersection kernel. The penalty parameter is selected by 10-fold cross-validation in the training set, whereas the scale parameter of the RBF kernel is selected as the mean χ^2 distance of the training samples. The RBF-χ^2 SVM achieves an improvement of $1-2\%$ over the intersection one.

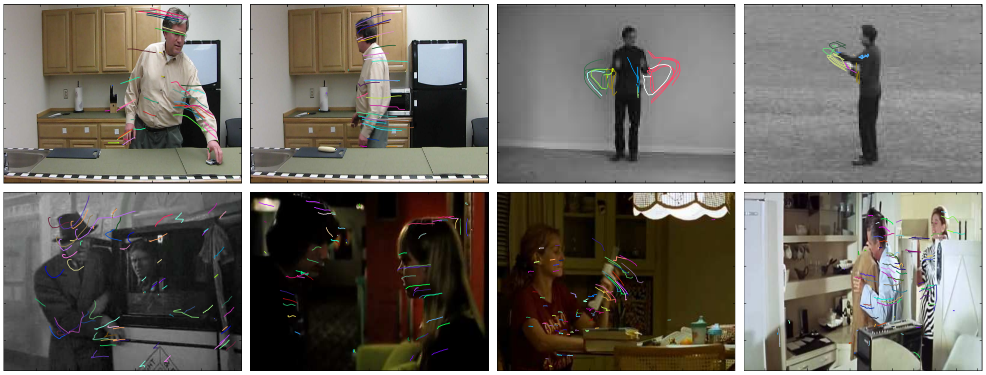

Fig. 1. *Tracks extracted from ADL, KTH and HOHA datasets. Color indicates their label based to the tracklet descriptor dictionary.*

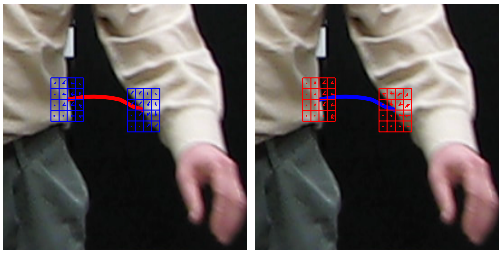

Fig. 2. *A track with samples of the histogram of gradient orientation (left, blue) and histogram of optical flow (right, red) along the trajectory. These are concatenated to form a 256-dimensional temporal sample of the time series that represents that elementary action.*

4 Experimental Evaluation

We evaluate the proposed scheme on three publicly available datasets: KTH [34] Activities of Daily Living (ADL) [27] and Hollywood Human Actions (HOHA) [17]. As pointed out in Sect. 3.2, AoG cannot be simply concatenated with either AoF or HoF, but has to be combined using multiple kernels. In our first two experiments we use the compact tracklet descriptor based on the HoG/HoF, so we can use one dictionary and one kernel, and have a fair comparison with existing local descriptors [7]. In the most challenging dataset (HOHA) the individual components HoG and AoG are compared in Table 3, and their combination with HoF is reported in Table 4.

KTH is chosen because of its popularity, though its modest spatial (160 × 120 pixels) and temporal (25 frames per second) resolution make for an impoverished data stream that is not well suited for local representations. There are 6 actions performed by 25 subjects in 4 scenarios (outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoor (s4)), resulting in 598 clips. The simplicity of these actions, combined with an uncluttered static background, make this dataset ideally suited for global representations [20]. Nevertheless, even without exploiting background subtraction or the global evolution of the silhouette (hard to obtain in most realistic scenarios), our scheme is competitive with the state of the art (Table 1).

More specifically, we track an average of 340 trajectories per video with an average length $\hat{T}_i = 23$ frames. Low resolution and the compression artifacts are a challenge to tracking, so the average length is relatively small. Our base regions D_i are fixed at 18×18 pixels, similar to the spatial size of Cuboids [7, 28, 29]. Examples of tracks and the corresponding HoG descriptors are shown in Fig. 3. The classification performance of algorithms that use spatio-temporal descriptors computed in volumes around interest points [21, 17, 4, 29, 7, 28] has proven that the choice of the temporal scale is crucial. Laptev *et al.* [17] construct static HoG/HoF around points detected by spatio-temporal Harris-3D [16] at multiple scales, using $\Delta t = 25, 36$; [4] computes a HoG/HoF around points detected by [23] in a volume with $\Delta t = 60$. Instead, our descriptors have variable temporal length depending on the image region D_i. Moreover, the optical flow in the image regions D_i can be estimated reliably. This is not the case for the spatio-temporal cubes around a specific interest point.

We use leave-one(person)-out cross validation and average the results over the 25 permutations. To construct the codebook we use a relatively small training set, similar to [28], to examine the generalization of our algorithm. We only use the descriptors extracted from the first two parts of the 72 videos of 3 subjects. Those descriptors are excluded from the test and training sets. It should be noted that [21, 4] used the videos of 24 subjects to construct the codebook, whereas [17] used 8 subjects. Using a codebook with 1560 TDs of HoG/HoF, we achieve 94.5% recognition rate using RBF-χ^2 SVM (Table 1) considering the dataset as a single large set (all average in one). Using linear SVM with intersection kernel we achieve 93.82% recognition rate. Considering each scenario separately the recognition rate is : (s1) 98%, (s2) 92.67%, (s3) 91.95% , (s4) 96.67%.

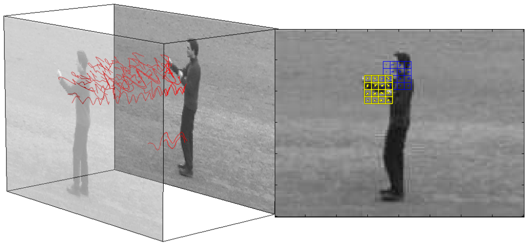

Fig. 3. *Example of the tracks and an instance of the corresponding appearance descriptor of a boxing action on the KTH dataset.*

We could push the performance of our algorithm by optimizing the weights between the different components of the features (spatial, motion), but our point is not to propose an action recognition system, but just to evaluate descriptors, so we refrain.

The **ADL** dataset has higher-resolution (1280 × 720 pixels at 30FPS) with 10 different complex activities targeted to an assisted living scenario (e.g. "an-

	evaluation	Recognition Rate		Structural Information
		all scenarios in one	average of all scenarios	
Our tracklets	Leave-One-Out	**94.5**%	**94.8**%	No
Niebles *et al.* [28]	Leave-One-Out	81.5%	N/A	No
Dollár *et al.* [7]	Leave-One-Out	81.2%	N/A	No
Schuldt *et al.* [34]	Split	71.7%	N/A	No
Nowozin *et al.* [29]	Split	84.7%	N/A	No
Liu *et al.* [22]	Leave-One-Out	N/A	94.15%	Yes
Lin *et al.* [20]	Leave-One-Out	93.4%	**95.8**%	Yes
Messing *et al.* [27]	Split	74%	N/A	No
Yao *et al.* [41]	Split	87.8%	N/A	Yes
Laptev *et al.* [17]	Split	91.8%	N/A	Yes
Jhuang *et al.* [11]	Split	N/A	91.7%	No
Schindler *et al.* [33]	Split	92.7%	90.7%	No
Yeffet *et al.* [42]	Split	90.1%	N/A	Yes
Chen *et al.* [4]	Leave-One-Out	95.0%	N/A	No

Table 1. *Performance comparison on KTH dataset. Despite not using background subtraction or structural information, our approach is competitive with the state of the art.*

swering phone (aP),” “eating snack (eS),” “eating banana (eB)”). Five subjects perform each activity thrice for a total of 150 clips of duration varying between 10 and 60 seconds. It has drawbacks similar to KTH, in that all actions are taken against a still background from a fixed vantage point, an incentive to overfitting by using background subtraction and global statistics such as the absolute position of tracks in the image. Despite not using absolute positions, a simple classifier based on TDs HoG/HoF outperforms the state of the art by a sizeable margin. We extract on average 1300 tracks with mean duration $\hat{T}_i = 110$ frames. The base regions D_i are fixed at 36×36 pixels. We again use leave-one (person)-out evaluation, similar to [27, 26], and report the average over the 5 permutations of the dataset. We randomly sampled $25K$ tracklets from the training set and constructed a dictionary with 2900 elements. Using this dictionary we achieve 82.67% average recognition rate using RBF-χ^2 SVM (Table 2). Comparison to [27] shows that our tracklet descriptor achieves comparable results without using any structural information (relative position or absolute position). It outperforms [27] even when their classifier uses the position of the extracted trajectories relative to the position of the face of the actor. In order to have a fair comparison with existing methods that report results in the ADL dataset, we incorporate a codebook of the absolute position $(\bar{g}_i(t), \bar{t})$ of the tracks with size 60 obtained using K-means. Combining linearly the two χ^2 kernels, we achieved 90% average recognition rate. We should note that, although absolute position is relevant in this dataset, and in particular it helps boost the performance of our algorithm as well as [27] significantly, it does so only because all sequences are taken from the same vantage point, in an environment with fixed layout. In general, we advocate *not* using absolute position, even if it improves the performance in this particular dataset.

The **HOHA** dataset overcomes the limitations of ADL and KTH. The dataset contains 430 movie videos (240×450 at 24FPS) with challenging camera motion, rapid scene changes and cluttered and unconstrained background. Moreover, the

	Recognition Rate
Our Tracklets	**82.67**%
Spatio-temporal cuboids [7] (implemented by [26])	43%
Velocity Histories [27]	63%
Latent Velocity Histories [27]	67%
Augmented Velocity Histories with Relative Position [26]	**72**%
Augmented Velocity Histories with Relative and Absolute Position [27]	89%

Table 2. *Performance comparison on ADL. Despite not using structural information or background subtraction, we improve the state of the art by a large margin. Using structural information, which we do not advocate, we can further improve recognition rate to* 90%*, highlighting the limitations of this particular dataset.*

human actions that are included are not constrained to single actor behaviors, e.g. "Sit down", but also interactions between humans, e.g. "Kiss", and objects, e.g. "Get Out of a Car". We evaluate our trajectory descriptors following the experimental setting proposed by [17], i.e. the test set has 211 videos with 217 labels and the training set has 219 videos with 231 labels (manually annotated). For each action we train a binary classifier and we evaluate our performance with average precision (AP) of the precision/recall curve.

In order to manage the large variability of the image sequences contained in the dataset, features [35] are detected in multiple scales. We extract on average 500 tracks with mean duration $\hat{T}_i = 51$ frames. For each image region D_i a HoG, HoF and AoG descriptor is constructed as described in (Sect. 3). First, a dictionary is created for each individual component of our tracklet descriptors and we evaluate its performance using RBF-χ^2 SVM (Table 3). Our TD of optical flow significantly outperforms the HoF proposed by Laptev *et al.* [17], proving to be more robust to background motion and large viewpoint changes. We also note that the performance of TD HoF is slightly worse than the trajectory transition descriptor (TTD) [38], which is combined with spatio-temporal grid to incorporate some structural information in the descriptor. Our TD of AoG outperforms marginally both our TD HoG and the HoG of [17], at a significantly reduced computational cost. Next, we construct our compact HoG/HoF tracklet descriptor and with a codebook with 2220 elements we achieve 32.1% mean average precision (MAP) (Table 4). In order to fuse the TD AoG feature descriptor with TD HoF feature in our classification framework, we build a kernel as a convex combination of their χ^2 kernels: $K_{AoG-HoG} = \lambda K_{AoG} + (1-\lambda) K_{HoF}$, λ was selected using cross-validation in the training set. The performance of the obtained kernel is 34.3% MAP. Our TD descriptors outperforms all the local descriptors that have been evaluated in HOHA dataset in a bag-of-features setting [14, 25, 17] and we are competitive with the holistic approach proposed by [42] and the methods that use multi-channel Gaussian kernels [17, 38] for combining the 48 or more channels provided by spatio-temporal grids.

5 Discussion

We have presented local spatio-temporal descriptors intended as low-level statistics to be used in action recognition systems. Our descriptors are deduced from an explicit model with all assumptions explicitly stated. They do not involve top-down modeling and can be efficiently learned from data. They can capture

Class	Our Tracklet			Laptev *et al.* [17]	
	HoG BoF	HoF BoF	AoG BoF	HoG BoF	HoF BoF
Answer phone	24.9%	22.1%	**33**%	13.4%	**24.6**%
Get out of car	21.1%	**19.3**%	**22.3**%	21.9%	14.9%
Hand shake	**20.4**%	**19.1**%	17.4%	18.6%	12.1%
Hug person	22.3%	**28.2**%	22.0%	**29.1**%	17.4%
Kiss	48.4%	**47.0**%	47.5%	**52.0**%	36.5%
Sit down	21.8%	**22.2**%	22.5%	**29.1**%	20.7%
Sit up	**16.7**%	**17.5**%	15.3%	6.5%	5.7%
Stand up	40.5%	**59.9**%	40.2%	**45.4**%	40.0%
MAP	27.1%	**29.4**%	**27.5**%	27.0%	21.5%

Table 3. *Performance comparison on HOHA Dataset of Individual components of Descriptors*

Class	Our Tracklet		Laptev *et al.* [17]		Yeffet *et al.* [42]	Matikainen *et al.* [25]	Kläser *et al.* [14]	Sun *et al.* [38]	
	HoG/HoF BoF	AoG-HoF BoF	Single	Combined		BoF	BoF	TTD Combined	TTD-SIFT Combined
Answer phone	26.7%	33.0%	26.7%	32.1%	35.1%	35.0%	18.6%	N/A	N/A
Get out of car	28.1%	27.0%	22.5%	41.5%	32.0%	7.7%	22.6%		
Hand shake	18.9%	20.1%	23.7%	32.3%	33.8%	5.3%	11.8%		
Hug person	25.0%	34.5%	34.9%	40.6%	28.3%	23.5%	19.8%		
Kiss	51.5%	53.7%	52.0%	53.3%	57.6%	42.9%	47.0%		
Sit down	23.8%	27.4%	37.8%	38.6%	36.2%	13.6%	32.5%		
Sit up	23.9%	19.0%	15.2%	18.2%	13.1%	11.1%	7.0%		
Stand up	59.1%	60.0%	45.4%	50.5%	58.3%	42.9%	38.0%		
MAP	32.1%	34.3%	32.9%	38.4%	36.8%	22.8%	24.7%	30.3%	44.94%

Table 4. *Performance comparison on HOHA Dataset.*

the discriminative statistics of the local causal structure of the data (temporal ordering), and the local shape and deformation of each base region. However, they do not enforce global shape or motion statistics, nor global temporal ordering. They could be used as a building block of more complex models for the recognition and classification of actions.

Although our goal is not to present a complete action recognition system, in order to test our descriptors we have employed them in simple classification schemes to recognize actions in commonly used benchmark datasets. In all cases, we obtain results comparable to or exceeding the state of the art, despite not making use of top-down structure.

Acknowledgments. This research is supported by ONR N00014-08-1-0414, AFOSR FA9550-09-1-0427, and ARO 56765-CI.

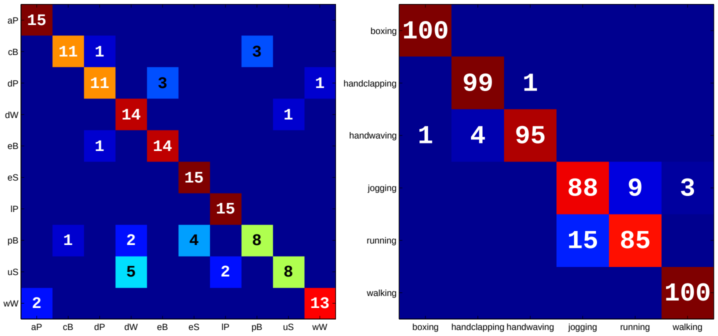

Fig. 4. *Confusion matrices for ADL dataset (**Left**) and for KTH dataset (**Right**)*

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. Lecture notes in computer science 3951, 404 (2006)
2. Birchfield, S.: Klt: An implementation of the kanade-lucas-tomasi feature tracker (1996)
3. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Trans. on Pattern Anal. and Machine Intell (2001)
4. Chen, M., Mummert, L., Pillai, P., Hauptmann, A., Sukthankar, R.: Exploiting multi-level parallelism for low-latency activity recognition in streaming video. In: Proc. of the first annual ACM SIGMM Conf. on Multimedia systems. ACM (2010)
5. Csurka, G., Dance, C.R., Dan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. of the Eur. Conf. on Computer Vision (ECCV) (2004)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2005)
7. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (October 2005)
8. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: Proc. Intl. Conf. on Computer Vision (2003)
9. Frey, B., Dueck, D.: Clustering by passing messages between data points. Science 315(5814), 972 (2007)
10. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey vision conference. vol. 15, p. 50. Manchester, UK (1988)
11. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Proc. Intl. Conf. on Computer Vision (2007)
12. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perceiving events and objects (1973)
13. Kaâniche, M., Brémond, F.: Gesture recognition by learning local motion signatures. In: Proc. Conf. Computer Vision and Pattern Recognition (2010)
14. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: British Machine Vision Conference. pp. 995–1004 (sep 2008)
15. Kumar, M., Patel, N., Woo, J.: Clustering seasonality patterns in the presence of errors. In: Proceedings of the eighth ACM SIGKDD (2002)
16. Laptev, I.: On space-time interest points. Intl. J. of Comp. Vis. 64(2), 107–123 (2005)
17. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. Conf. Computer Vision and Pattern Recognition (2008)
18. Laptev, I., Pérez, P.: Retrieving actions in movies. In: Proc. Intl. Conf. on Computer Vision (2007)
19. Lee, T., Soatto, S.: An end-to-end visual recognition system. Technical Report UCLA-CSD-100008 (February 10, 2010, revised March 18, 2010)
20. Lin, Z., Jiang, Z., Davis, L.: Recognizing actions by shape-motion prototype trees. In: Proc. Intl. Conf. on Computer Vision (2009)
21. Liu, J., Luo, J., Shah, M.: Recognizing Realistic Actions from Videos “in the Wild”. In: Proc. IEEE Computer Vision and Pattern Recognition (2009)
22. Liu, J., Shah, M.: Learning human actions via information maximization. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2008)

23. Lowe, D.: Object recognition from local scale-invariant features. In: Intl. Conf. on Computer Vision (1999)
24. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. 7th Int. Joint Conf. on Art. Intell. (1981)
25. Matikainen, P., Hebert, M., Sukthankar, R.: Trajectons: Action recognition through the motion analysis of tracked features. In: ICCV workshop on Video-oriented Objected and Event Classification (2009)
26. Messing, R., Pal, C.: Behavior recognition in video with extended models of feature velocity dynamics. In: AAAI Spring Symposium Technical Report (2009)
27. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: Intl. Conf. on Computer Vision (2009)
28. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. Intl. J. of Comp. Vis. 79(3) (2008)
29. Nowozin, S., Bakir, G., Tsuda, K.: Discriminative subsequence mining for action classification. In: Proc. Intl. Conf. on Computer Vision (2007)
30. Rabiner, L., Juang, B.: Fundamentals of speech recognition. Prentice hall (1993)
31. Robert, C.P.: The Bayesian Choice. Springer Verlag, New York (2001)
32. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26(1), 43–49 (1978)
33. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2008)
34. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: Proc. Intl. Conf. on Pattern Recognition (2004)
35. Shi, J., Tomasi, C.: Good features to track. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (1994)
36. Soatto, S.: Towards a mathematical theory of visual information (2010)
37. Soatto, S., Yezzi, A.: Deformotion: deforming motion, shape average and the joint segmentation and registration of images. In: Proc. of the Eur. Conf. on Computer Vision (ECCV). vol. 3, pp. 32–47 (2002)
38. Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2009)
39. Veeraraghavan, A., Chellappa, R., Roy-Chowdhury, A.: The function space of an activity. In: Proc. IEEE Conf. On Computer Vision and Pattern Recognitio (2006)
40. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference (2009)
41. Yao, B., Zhu, S.: Learning Deformable Action Templates from Cluttered Videos. Intl. Conf. on Computer Vision (2009)
42. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: Proc. Intl. Conf. on Computer Vision (2009)
43. Zelnik-Manor, L., Irani, M.: Statistical analysis of dynamic actions. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1530–1535 (2006)