Dynamically Encoded Actions based on Spacetime Saliency

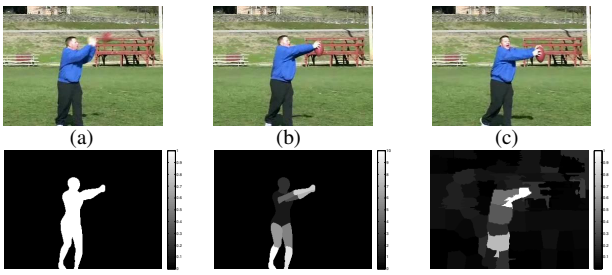
Christoph Feichtenhofer[1] Axel Pinz[1] Richard P. Wildes[2]
[1]Institute of Electrical Measurement and Measurement Signal Processing, TU Graz, Austria
[2]Department of Electrical Engineering and Computer Science, York University, Toronto, Canada
{feichtenhofer, axel.pinz}@tugraz.at wildes@cse.yorku.ca

Abstract

Human actions typically occur over a well localized extent in both space and time. Similarly, as typically captured in video, human actions have small spatiotemporal support in image space. This paper capitalizes on these observations by weighting feature pooling for action recognition over those areas within a video where actions are most likely to occur. To enable this operation, we define a novel measure of spacetime saliency. The measure relies on two observations regarding foreground motion of human actors: They typically exhibit motion that contrasts with that of their surrounding region and they are spatially compact. By using the resulting definition of saliency during feature pooling we show that action recognition performance achieves state-of-the-art levels on three widely considered action recognition datasets. Our saliency weighted pooling can be applied to essentially any locally defined features and encodings thereof. Additionally, we demonstrate that inclusion of locally aggregated spatiotemporal energy features, which efficiently result as a by-product of the saliency computation, further boosts performance over reliance on standard action recognition features alone.

1. Introduction

Recently, a great amount of effort has addressed video-based action recognition, with many approaches employing the Bags-of-visual-Word (BoW) principle consisting of three general steps: *1)* In the *primitive feature extraction* step, local features are extracted either from interest points or densely from regular locations by application of hand-designed or learned filters. *2)* A *feature transformation* step generates intermediate representations that map local features into more effective representations for the underlying task (*e.g.* by using unsupervised or supervised learned visual words). *3)* A *pooling* step accumulates transformed features over pre-defined regions. Steps *2)* and *3)* may be applied several times in a hierarchical manner (as *e.g.* in deep learning methods [20, 23, 37]). Finally, these pooled, transformed features are fed into a classifier for decision.



(a) (b) (c) (d) GT-mask (e) GT-flow (f) $\mathcal{S}_T$

Figure 1. Our proposed measure of spacetime saliency to enhance feature pooling for action recognition. (a)-(c) Input frames of a video from the J-HMDB dataset [19] showing a catch action. (d) Ground truth puppet mask annotation [19]. (e) The puppet flow generated from ground truth body part annotations [55]. (f) Our spacetime saliency measure, $\mathcal{S}_T$, for weighting the contribution of local spatiotemporal features for action recognition. Note the large similarities between the user-annotated puppet flow and our saliency measure which is computed efficiently from motion statistics without any form of supervision.

For action recognition, typical techniques used in the BoW steps are: *1)* SIFT [28], HOG3D [21], spacetime correlation patches [36], Histograms of Optical Flow (HOF) [25], Motion Boundary Histograms (MBH) [12], trajectories [42], and Spatiotemporal Oriented Energy (SOE) [13]; *2)* Locality-constrained Linear Coding (LLC) [45], Super Vector (SV) [52], Vector of Locally Aggregated Descriptors (VLAD) [17] and Fisher Vectors (FV) [32]; *3)* average- and max-pooling with geometry embedded by aggregating with Spatial Pyramid Matching (SPM) [27, 50]. Classification is mostly realized using a Support Vector Machine (SVM) [9].

Pooling typically has been performed across an entire video, sometimes in conjunction with spatial pyramids [27]. Some research, however, has investigated selective pooling over regions where actors are likely to be present. From the perspective of what might be accomplished in principle, research has shown that hand annotated ground truth indicating where actors are present can improve action recognition [19] (see *e.g.* Figure 1(d) and 1(e)). From a more practical perspective, various approaches have been developed

for automatic recovery of areas where actors are likely to be present. Some of this research relies on explicit detection of human actors or objects to be acted upon [34, 40]. Other research has instead relied on grouping various dynamic measurements (*e.g.* dense trajectories [35], SOEs [46], tracklets [14]) with subsequent pooling restricted to the resulting groups. Still other research has more abstractly defined and realized a measure of actionness [6] by analogy with that of objectness employed in conjunction with object detection [8]; however, that work did not explicitly embed its measure in a complete action recognition algorithm. Somewhat differently, research has sought to discard feature codes that do not arise from action regions via a two-stage process of image stabilization and suppression of non-moving edges [22]. Other less related work, as it relies on human performance data, used eye tracking to constrain action recognition [41].

The most closely related work to ours is previous research that has made explicit use of notions of saliency to capture foreground regions where actions are most likely to occur. One such approach [3] made use of three types of saliency measures to pool dense trajectory features [42]. In distinction from our work, they employ three very simple saliency measures (detected corners, image brightness and motion magnitude) and focus their work on the learning component of the system, where they propose a weighted SVM classifier. Also related is work that combines colour and motion gradients via a graph-based saliency measure to select foreground actions [39]. Their approach relies on computation of optical flow, while ours instead relies on spatiotemporal oriented filtering, which can capture a wider range of dynamic patterns (*e.g.* multiple motions in a region of analysis where action parts overlap, and temporal flicker [13]) and also is less computationally expensive. Their computational expense is further increased in comparison to ours by use of a 3D MRF optimization process to generate foreground weights.

In this work we tackle the problem of selective pooling from a different direction. We present a novel approach to action recognition, based on spacetime saliency, as illustrated in Fig. 1, which shows a sequence of a person catching a ball as our running example. Our approach dynamically encodes and pools primitive feature measurements via a new definition of spacetime saliency weights based on directional motion energy contrast and spatial variance to capture actions. These complementary weights allow the approach to reap the benefit of pooling over regions of likely action occurrence so that recognition is uncorrupted by irrelevant or distracting data. In general, the encoding and pooling approach can be applied essentially to any local feature measurements; here it is illustrated using various combinations of improved dense trajectories [43] and a novel extension to SOEs. Significantly, in empirical evaluation the approach is competive with and can even exceed the previous state-of-the-art in action recognition on three standard datasets, including J-HMDB [19], HMDB51 [24] and UCF101 [38].

2. Spacetime saliency-based feature encoding

This section documents the proposed approach to encoding and pooling of primitive features to capture actions. Since the approach is largely applicable to any local feature measurements, it initially is cast in terms of arbitrary feature vectors, $\mathbf{f}(\mathbf{x})$, with $\mathbf{x} = (x, y, t)^\top$ image spacetime coordinates. The essential notion is to define a local measure of foreground motion saliency, $\mathcal{S}(\mathbf{x})$, that is highest in regions likely to contain an action. In the remainder of this section, we begin by defining our local measure of motion in terms of spatiotemporal oriented energy filtering operations. With these measurements in hand, we define our measure of saliency, $\mathcal{S}(\mathbf{x})$. Finally, we show how to use this measure to dynamically encode and pool features, $\mathbf{f}(\mathbf{x})$, to capture actions.

2.1. Directional motion energy

There are three steps to the proposed approach to capturing directional motion energy as the measurements over which saliency is defined. The first step involves operating on input video imagery with filters tuned for various spatiotemporal orientations. The second step combines the raw filter outputs to capture a specific set of motion directions. The third step normalizes the combined filter outputs for better photometric invariance.

The first step is realized by convolving the video with 3D Gaussian third derivative filters, $G_{3D}^{(3)}(\theta_i, \sigma_j) = \kappa \frac{\partial^3}{\partial \theta_i^3} \exp\left(-\frac{x^2+y^2+t^2}{2\sigma_j^2}\right)$, with θ_i and σ_j denoting the 3D filter orientations and scales, respectively, and κ providing normalization to yield

$$E_{\text{ST}}(\mathbf{x}; \theta_i, \sigma_j) = |G_{3D}^{(3)}(\theta_i, \sigma_j) * \mathcal{V}(\mathbf{x})|^2, \qquad (1)$$

with the grayscale spacetime volume, $\mathcal{V}$, indexed by $\mathbf{x} = (x, y, t)^\top$, formed by stacking all video frames of a sequence along the temporal axis, t. Subscript ST on E_{ST} denotes *spatiotemporal* orientation.

The second step combines the spatiotemporal responses, (1), to yield measures of motion information along certain directions independent of spatial appearance, as follows. In the frequency domain, motion occurs as a plane through the origin [47]. Therefore, summation across a set of $x-y-t$ energy measurements consistent with a single frequency domain plane through the origin is indicative of the associated spacetime orientation, *e.g.* motion direction, independent of purely spatial orientation. Let the plane be defined by its normal, $\hat{\mathbf{n}} = (n_x, n_y, n_t)^\top$, then measurements of orienta-

tion consistent with this plane are given as

$$E_{\mathrm{T}}(\mathbf{x};\hat{\mathbf{n}}_k,\sigma_j)=\sum_{i=0}^{N}E_{\mathrm{ST}}(\mathbf{x};\theta_i,\sigma_j), \tag{2}$$

with θ_i one of $N+1$ equally spaced orientations consistent with the frequency domain plane and $N=3$ the order of the employed Gaussian derivative filters; for details see [13]. Here, the subscript T on E_{T} serves to denote that the spatiotemporal measurements have been "*marginalized*" with respect to purely spatial orientation. To further suppress noise and smooth the filter responses (2), they are blurred by a 5-tap Gaussian $G_{3D}^{(1)}$ for subsequent processing.

The derived measurements, (2), can be taken as providing measures of the motion energy along the specified directions, $\hat{\mathbf{n}}$. This interpretation is justified by Parseval's theorem [29], which states that the sum of the squared values over the spacetime domain is proportional to the sum of the squared magnitude of the Fourier components over the frequency domain. Thus, for every location, $\mathbf{x}$, the local motion energy $E_{\mathrm{T}}(\mathbf{x};\hat{\mathbf{n}}_k,\sigma_j)$ measures the power of local structure along each considered orientation $\hat{\mathbf{n}}_k$ and scale σ_j.

Owing to the bandpass nature of the Gaussian derivative filters, the spatiotemporal orientation measurements are invariant to additive photometric variations (*e.g.*, as might arise from local image brightness change in imaged scenes). To provide additional invariance to multiplicative photometric variations for the energy measurements, (2), each motion direction selective measurement is normalized with respect to the sum of all filter responses at that point as

$$\hat{E}_{\mathrm{T}}(\mathbf{x};\hat{\mathbf{n}}_k,\sigma_j)=\frac{E_{\mathrm{T}}(\mathbf{x};\hat{\mathbf{n}}_k,\sigma_j)}{\sum_{m=1}^{M}E_{\mathrm{T}}(\mathbf{x};\hat{\mathbf{n}}_m,\sigma_j)+\epsilon}, \tag{3}$$

where M denotes the number of orientation measurements considered, to yield a normalized set of measurements, $\hat{E}_{\mathrm{T}}$. Note that ϵ is a small constant added to the sum of the energies over all orientations. This bias operates as a noise floor and avoids numerical instabilities at low overall energies. To explicitly capture lack of oriented spacetime structure, another feature channel

$$\hat{E}_{\mathrm{T}}^{\epsilon}(\mathbf{x};\sigma_j)=\frac{\epsilon}{\sum_{m=1}^{M}E_{\mathrm{T}}(\mathbf{x};\hat{\mathbf{n}}_m,\sigma_j)+\epsilon}, \tag{4}$$

is added to the contrast-normalized filter responses, (3). Note, *e.g.*, that regions lacking oriented structure will have the summation in (4) evaluate to 0; hence, $\hat{E}_{\mathrm{T}}^{\epsilon}$ will tend to 1 and thereby indicate relative lack of structure.

Our motion energy decomposition of a video showing a person catching a ball are shown in Fig. 2. The various panels depict 1 pixel/frame motion energy along the rightward $\hat{\mathbf{n}}_r=(1,0,1)^{\top}$, leftward $\hat{\mathbf{n}}_l=(-1,0,1)^{\top}$, upward $\hat{\mathbf{n}}_u=(0,1,1)^{\top}$ and downward $\hat{\mathbf{n}}_d=(0,-1,1)^{\top}$ directions as well as horizontal and vertical flicker $\hat{\mathbf{n}}_{hf}=(1,0,0)^{\top}$ and $\hat{\mathbf{n}}_{vf}=(0,1,0)^{\top}$ (infinite velocity), static $\hat{\mathbf{n}}_s=(0,0,1)^{\top}$ (zero velocity) and unstructured energy.

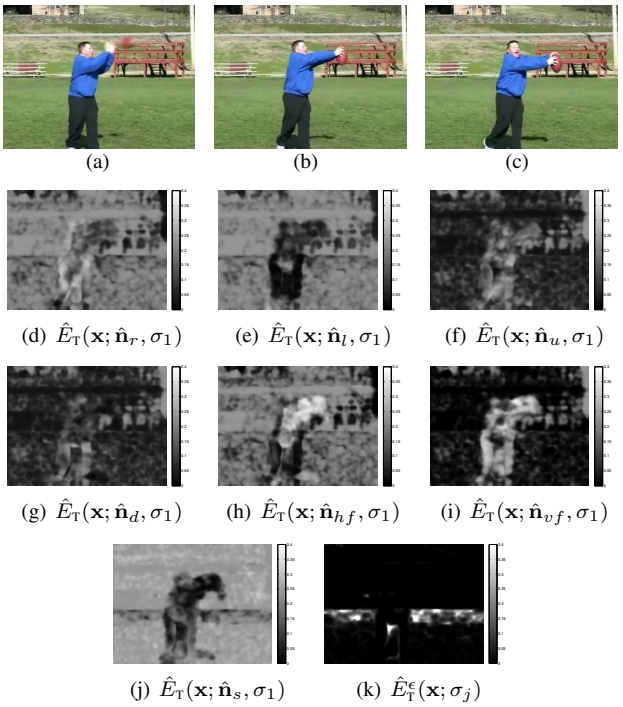


Figure 2. 10th (a), 15th (b) and 20th (c) frame of a video from the HMDB dataset [24], showing a catch action. Normalized energies (3) for the 15th frame are shown in (d)-(j) for various frequency planes $\hat{\mathbf{n}}$: rightward (d), leftward (e), downward (g), upward (f), horizontal flicker (h), vertical flicker (i) and static (j). Further, (k) illustrates the no structure channel, (4).

2.2. Spacetime saliency

Our essential notion is to define a local measure of saliency, $\mathcal{S}(\mathbf{x})$, that is highest in regions likely to contain an action. To define this measure we rely on two general observations regarding actions. First, actions typically involve a foreground motion that is distinct from the surrounding background. Indeed, even in the presence of global camera motion, a foreground action will exhibit a different (superimposed) pattern of motion. For example, a participant in a sporting event will yield a motion that is distinct from that of overall camera motion when he or she is engaged in their sporting activity. We refer to this property as *motion contrast*. Second, action patterns typically are spatially compact, while background motions are more widely distributed across an image sequence. For example, even interactions between two people (*e.g.* a hug, handshake or kiss) occupy a relatively small portion of an image. We refer to this second property as *motion variance*. In combination, these two properties are used to define our measure of spacetime saliency, $\mathcal{S}(\mathbf{x})$, for capturing foreground action motion. While our notions of saliency being defined in terms of

both local contrast and global variance are present in other definitions of image saliency (*e.g.*, [7, 15, 31, 49, 54]), it appears that our approach is the first to instantiate them in terms of motion measurements for action recognition.

2.2.1 Motion contrast

As a preliminary step to computing motion contrast, we perform a coarse segmentation of the imagery to obtain regions between which contrast is defined. In the current implementation, this segmentation is performed in terms of SLIC superpixels [1], setting the average number of pixels in each superpixel to 1000. Figures 3(a)-3(c) illustrate the segmentations for the ball catching sequence shown earlier in Figure 2. To represent the temporal properties of each spatial element i, we average the normalized energies within it

$$\hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}, \sigma_j) = \frac{1}{\|\Omega_i\|} \sum_{\mathbf{x} \in \Omega_i} \hat{E}_{\mathrm{T}}(\mathbf{x}; \hat{\mathbf{n}}, \sigma_j), \tag{5}$$

with Ω_i the spatial support of superpixel i. In performing the energy aggregations, (5), since a region without structure cannot be distinguished from a static region, we combine the static and unstructured channels of the elements

$$\hat{E}_{\mathrm{T}}^{(i,s+\epsilon)} = \hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}_s, \sigma_j) + \hat{E}_{\mathrm{T}}^{(i,\epsilon)}(\sigma_j) \tag{6}$$

An example for the resulting energy distribution is shown in Figures 3(d)-3(j).

Given two elements, i and j, their motion contrast is calculated in terms of the difference between their motion characteristics, *i.e.* distributions of directional motion energies, (3), weighted by their spatial distance. Difference in motion characteristics is given in terms of the Hellinger distance (also known as the Bhattacharyya distance) [2],

$$d_{\mathrm{H}}^{(i,j)} = \left|\left| \sqrt{\hat{E}_{\mathrm{T}}^{(i)}} - \sqrt{\hat{E}_{\mathrm{T}}^{(j)}} \right|\right|_2. \tag{7}$$

We choose the Hellinger distance, since the directional motion energies, $\hat{E}_{\mathrm{T}}$, represent ℓ_1 normalized distributions, which typically are compared more effectively by using histogram distance measures, compared to *e.g.* the Euclidean distance. Note that the Hellinger distance can simply be computed by element-wise square-rooting before computing the Euclidean distance [2].

Spatial distance plays into the saliency calculation via multiplicative weighting with an exponential on the Euclidean distance between the centre of mass coordinates of the elements, i and j, $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$, respectively,

$$e^{-||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}||_2^2}. \tag{8}$$

Finally, to obtain an overall measure of contrast for element i, its pairwise contrast with all other elements, j, is summed

$$\mathcal{S}_{\mathrm{T,CTR}}^{(i)} = \sum_{j=1}^{J} d_{\mathrm{H}}^{(i,j)} e^{-||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}||_2^2}, \tag{9}$$

with J the total number of elements, *i.e.* superpixels. Example contrast measurements, $\mathcal{S}_{\mathrm{T,CTR}}^{(i)}$, for the ball catching sequence are shown in Figure 3(k).

2.2.2 Spatial motion variance

The second measure of saliency is based on the common observation that foreground motion typically occurs in a spatially localized region of a video, whereas background motion (including no motion, *i.e.* static) is typically widely distributed over the video. Our second saliency measure therefore ranks local regions as highly salient if the spatial variance of their motion characteristics is low.

We define spatial variance of motion by analogy with the standard definition of the variance of a (discrete random) variable, x^i, with probability mass function, $p(x^i)$, *i.e.* as $\sum_{i=1}^{n} p(x^i) \times (x^i - \mu)^2$ with $\mu = \sum_{i=1}^{n} p(x^i) \times x^i$. In our analogy, we let $x^i = \mathbf{x}^{(i)}$ be the centroid coordinates of segment i and $p(x^i) = e^{-d_{\mathrm{H}}^{(i,j)}}$, *i.e.* an exponential in the difference between the motion characteristics of elements i and j, (7). Correspondingly, we define the correlate of the expected value, μ, to be the weighted average position of energy $\hat{E}_{\mathrm{T}}^{(i)}$, (5), *i.e.*

$$\bar{E}_{\mathrm{T}}{}^{(i)} = \gamma_i \sum_{j=1}^{J} x^{(j)} e^{-d_{\mathrm{H}}^{(i,j)}}, \tag{10}$$

with weights $e^{-d_{\mathrm{H}}^{(i,j)}}$ by direct analogy with the standard definition of μ, J taken so the summation ranges over all superpixels and $\gamma_i = \frac{1}{\sum_{j=1}^{J} e^{-d_{\mathrm{H}}^{(i,j)}}}$ necessitated by the need for normalization. Thus, the definition of spatial variance of motion of segement i becomes

$$E_{\mathrm{VAR}}^{(i)} = \sum_{j=1}^{J} ||x^{(j)} - \bar{E}_{\mathrm{T}}{}^{(i)}||_2^2 e^{-d_{\mathrm{H}}^{(i,j)}}. \tag{11}$$

Our measure of spatial variance of motion, (11), increases as the the motion of i becomes more spread out across an image, while for saliency we seek the opposite. Given that $E_{\mathrm{VAR}}^{(i)}$ is normalized by construction, we arrive at the desired measure simply by defining

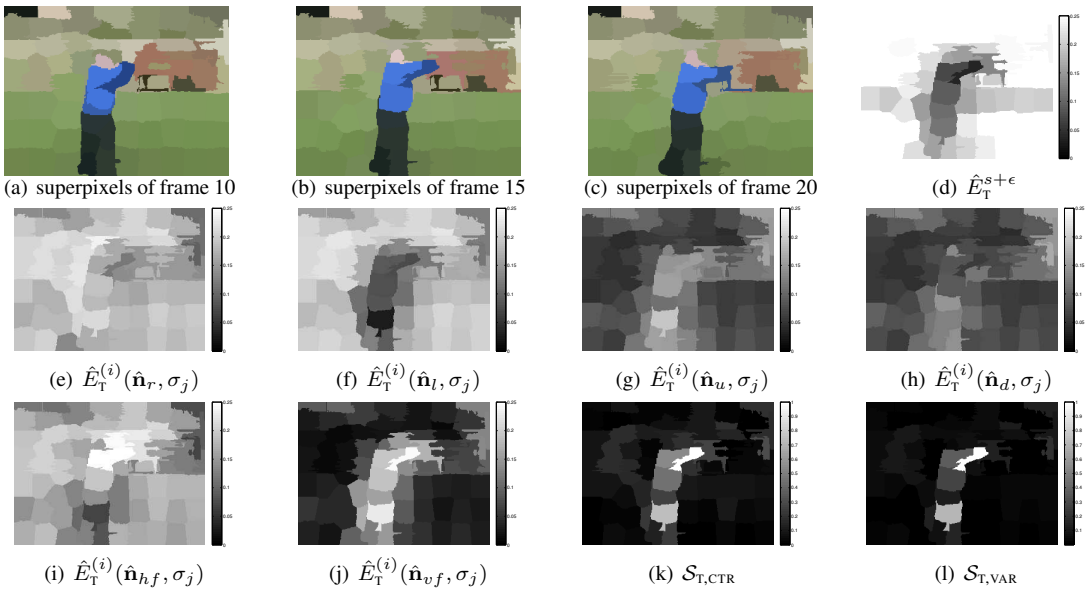
$$\mathcal{S}_{\mathrm{T,VAR}}^{(i)} = 1 - E_{\mathrm{VAR}}^{(i)}. \tag{12}$$

Example variance measurements, $\mathcal{S}_{\mathrm{T,VAR}}^{(i)}$, for the ball catching sequence are shown in Figure 3(l).

Finally, the overall spacetime saliency is given by combining the two saliency measures so far defined, to yield

$$\mathcal{S}_{\mathrm{T}}^{(i)} = \frac{\mathcal{S}_{\mathrm{T,CTR}}^{(i)} + \mathcal{S}_{\mathrm{T,VAR}}^{(i)}}{2}. \tag{13}$$

Example overall saliency measurements, $\mathcal{S}_{\mathrm{T}}^{(i)}$, for the ball catching sequence are shown in Figure 1(f); additional examples are provided on the project webpage.



(a) superpixels of frame 10 (b) superpixels of frame 15 (c) superpixels of frame 20 (d) $\hat{E}_{\mathrm{T}}^{s+\epsilon}$

(e) $\hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}_r, \sigma_j)$ (f) $\hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}_l, \sigma_j)$ (g) $\hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}_u, \sigma_j)$ (h) $\hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}_d, \sigma_j)$

(i) $\hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}_{hf}, \sigma_j)$ (j) $\hat{E}_{\mathrm{T}}^{(i)}(\hat{\mathbf{n}}_{vf}, \sigma_j)$ (k) $\mathcal{S}_{\mathrm{T,CTR}}$ (l) $\mathcal{S}_{\mathrm{T,VAR}}$

Figure 3. Mean RGB colours of SLIC superpixels for the 10th (a), 15th (b) and 20th (c) frame of a catch sequence from the HMDB dataset; original frames shown in Fig. 2. The distribution of energies summed over superpixels are shown in (d)-(j). Image (d) shows the combination of static and unstructured energy via summation; images (e)-(j) show mean oriented energies in the superpixels across various directions, $\hat{\mathbf{n}}$: rightward (e), leftward (f), downward (h), upward (g), horizontal flicker (h), and vertical flicker (i). Images (k) and (l) illustrate our resulting computations of motion contrast (9) and spatial motion variance (12), respectively.

2.3. Dynamic feature encoding via saliency

Our spacetime saliency measure, $\mathcal{S}_{\mathrm{T}}^{(i)}$, can be used for weighted pooling with any encoding method in the BoW framework to enhance the contribution of local features $\mathbf{f}(\mathbf{x})$ by their saliency $\mathcal{S}_{\mathrm{T}}(\mathbf{x})$. Here, $\mathcal{S}_{\mathrm{T}}(\mathbf{x})$ is derived directly from $\mathcal{S}_{\mathrm{T}}^{(i)}$ by having the saliency of $\mathbf{x}$ be defined as that of the superpixel element i to which it belongs, *i.e.*

$$\mathcal{S}_{\mathrm{T}}(\mathbf{x}) = \mathcal{S}_{\mathrm{T}}^{(i)}; \quad \mathbf{x} \in \Omega_i, \tag{14}$$

with Ω_i the support of superpixel i, as before. Hence, features from spatial regions that likely correspond to the foreground motion of an action will have higher importance in the encoding procedure. As a concrete example, we illustrate use with Fisher vector (FV) encoding, as it provides the current state-of-the-art in video action recognition (*e.g.* [4, 16, 26, 43, 48, 51]).

FVs [32, 33] capture the gradient of the log-likelihood of the features, $\mathbf{f}(\mathbf{x})$, with respect to the parameters of a generative model. The model is learned on training descriptors via a Gaussian Mixture Model, (GMM), $p(\mathbf{f}(\mathbf{x})|\boldsymbol{\theta})$, with parameters $\boldsymbol{\theta} = (w_k, \mu_k, \sigma_k)^\top, k = 1, \ldots, K$, with w_k being the weight, μ_k the mean, and σ_k the diagonal covariance of the k^{th} mixture. Training of the GMM parameters is realized with the expectation maximization algorithm. For each of the $k = 1, \ldots, K$ mixtures, the soft assignment for a feature vector $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^D$ is denoted by $p(k|\mathbf{f}(\mathbf{x}), \boldsymbol{\theta})$.

$$p(k|\mathbf{f}(\mathbf{x}), \boldsymbol{\theta}) = \frac{w_k p_k(\mathbf{f}(\mathbf{x})|\mu_k, \sigma_k)}{\sum_{j=1}^{K} w_j p_j(\mathbf{f}(\mathbf{x})|\mu_j, \sigma_j)}. \tag{15}$$

An FV models mean and covariance gradients between features $\{\mathbf{f}_l(\mathbf{x})\}_{l=1}^{L}$ and the GMM modelled distribution

$$\mathbf{c}^{(\mathrm{FV})} = [\mathbf{c}_1^{(\mu)}, \mathbf{c}_1^{(\sigma)}, \ldots, \mathbf{c}_K^{(\mu)}, \mathbf{c}_K^{(\sigma)}]. \tag{16}$$

To apply our saliency measure, $\mathcal{S}_{\mathrm{T}}(\mathbf{x})$, to FVs, we employ it as a local weighting function during aggregation of the first- and second-order gradients:

$$\mathbf{c}_k^{(\mu)} = \frac{1}{N\sqrt{w_k}} \sum_{l=1}^{L} p(k|\mathbf{f}_l(\mathbf{x}), \boldsymbol{\theta}) \left(\frac{\mathbf{f}_l(\mathbf{x}) - \mu_k}{\sigma_k} \right) \mathcal{S}_{\mathrm{T}}(\mathbf{x}), \tag{17}$$

and

$$\mathbf{c}_k^{(\sigma)} = \frac{1}{N\sqrt{2w_k}} \sum_{l=1}^{L} p(k|\mathbf{f}_l(\mathbf{x}), \boldsymbol{\theta}) \left(\frac{(\mathbf{f}_l(\mathbf{x}) - \mu_k)^2}{\sigma_k^2} - 1 \right) \mathcal{S}_{\mathrm{T}}(\mathbf{x}). \tag{18}$$

Thus, for each feature type, we calculate a FV that is explicitly weighted in favour of spacetime saliency where actions are most likely to occur.

For classification, each resulting Fisher vector is used to train one-vs-rest SVM classifiers. During recognition, each feature type is processed by its one-vs-rest SVM classifier to yield match scores for a test video. All SVM scores subsequently are combined via averaging to yield a late fusion classification according to the maximum score. While more sophisticated fusion strategies could be considered (*e.g.* determination of fusion weights via cross-validation on training data), such considerations are left for future research.

3. Primitive features

To emphasize the generality of the proposed approach, the previous section cast saliency weighted encoding and pooling in terms of of arbitrary features, $\mathbf{f}$. This section briefly documents the two kinds of features that are used in the empirical evaluation of the approach presented in this paper. The two classes of features are selected as they are instances of two types used in BoW-based action recognition, as presented in the introduction: motion-based features and spatiotemporal orientation features. Moreover, in empirical evaluation they will be shown to be complementary in improving recognition performance.

The employed motion-based features are Dense Trajectories (DT) [42] or Improved Dense Trajectories (IDT) [43]. The descriptors employed in conjunction with these features include HOG [11], HOF [25] and MBH [10] descriptors. Details of the involved computations are exactly as in the original DT and IDT work and are suppressed here in the interest of space.

The employed spatiotemporal orientation features are a novel extension to (appearance marginalized) Spatiotemporal Oriented Energies (SOEs), originally applied to action recognition elsewhere [13]. In particular, we make use of multiscale oriented filtering to enrich the descriptor (the previous approach used only a single scale) as well as different subregion aggregations and normalization, which were found to improve performance in preliminary experiments. The resulting representation is well suited to capturing actions, as it allows the relative configuration of primitive measurements within a neighborhood to be made explicit, *e.g.* as useful in capturing the relative movements of limbs or actors. We term these features Locally Aggregated Temporal Energies (LATE).

Given local energy measurements, (3) and (4), defined at each point, $\mathbf{x}$, we construct the LATE descriptor as follows. We sample the local measurements with a spatiotemporal stride of $\Delta\mathbf{x}$ by centering cuboids of size $r_x \times r_y \times r_t$ around $\mathbf{x}$. In order to capture the local neighborhood structure within each cuboid, these regions are then subdivided into $c_x \times c_y \times c_t$ sub-regions over which energy measurements are aggregated into histograms. Within each subregion the aggregation takes the form of a simple sum over the sub-region's support, $\Omega(\mathbf{x})$, to yield

$$E_{\text{AGG}}(\mathbf{x}; \hat{\mathbf{n}}_k, \sigma_j) = \sum_{\tilde{\mathbf{x}} \in \Omega(\mathbf{x})} \hat{E}_{\text{T}}(\tilde{\mathbf{x}}; \hat{\mathbf{n}}_k, \sigma_j). \qquad (19)$$

Next, for each sample point, $\mathbf{x}$, a feature vector, $\mathbf{f}_{\text{LATE}}(\mathbf{x})$, is constructed by stacking the energy measurements, (19), that were summed within each sub-region of the surrounding cuboid. Finally, we apply RootSIFT normalization [2] (*i.e.* square rooting each vector element) followed by ℓ_2 normalization to the feature vector. LATE features are extracted in this manner for several spatial scales, while leaving the temporal scale unchanged.

4. Implementation details

We densely extract multi-scale LATE features in a scale-space pyramid with $|\sigma| = 5$ different scales by downsampling the image (*i.e.* spatial dimensions only) by factors of $\sqrt{2}$. We use a spatiotemporal stride of $\Delta\mathbf{x} = (x, y, t)^\top = (8, 8, 16)^\top$ for dense feature extraction. The features are computed over regions of size $r_x \times r_y \times r_t = 16 \times 16 \times 16$ that are further divided into $c_x \times c_y \times c_t = 2 \times 2 \times 3$ sub-regions for aggregation. Using 8 energy filterings, parametrized by $\hat{\mathbf{n}}$ results in $D_{\text{LATE}} = 2 \times 2 \times 3 \times 8 = 96$ dimensional LATE descriptors. The 8 values of $\hat{\mathbf{n}}$ are chosen to correspond to rightward $\hat{\mathbf{n}}_r = (1, 0, 1)^\top$, leftward $\hat{\mathbf{n}}_l = (-1, 0, 1)^\top$, upward $\hat{\mathbf{n}}_u = (0, 1, 1)^\top$ and downward $\hat{\mathbf{n}}_d = (0, -1, 1)^\top$ motion as well as horizontal and vertical flicker $\hat{\mathbf{n}}_{hf} = (1, 0, 0)^\top$ and $\hat{\mathbf{n}}_{vf} = (0, 1, 0)^\top$ (infinite velocity motion), static $\hat{\mathbf{n}}_s = (0, 0, 1)^\top$ (zero velocity motion) and unstructured energy $\hat{E}^{\epsilon}_{\text{T}}(\mathbf{x}; \sigma_j)$, *e.g.* as illustrated in Fig. 2. All DT and IDT parameters are used as in [42, 43] and we use their publicly available code to extract the descriptors.

For encoding, a random subset of features consisting of 100.000 descriptors from the training data, are used to learn a visual vocabulary. A GMM with $K = 256$ mixtures is fit to each of the subsampled training descriptors (HOG, HOF, MBH and LATE). PCA whitening is applied to the raw descriptors to reduce their dimension by a factor of two. Data decorrelation via PCA also supports the diagonal covariance assumptions in the employed GMM [18]. Before training the GMM all features are augmented with their normalized x, y and t coordinates. After dynamic aggregation of the local descriptors' mean (17) and covariance (18) gradients, the FVs are signed square-rooted and ℓ_2 normalized. Square-rooting is applied twice, once to the raw encodings and once again after ℓ_2 normalization [5].

For training, all feature vectors extracted from the training set are used to train one-vs-rest linear SVM classifiers. The SVM's regularization loss trade-off parameter is set to $C = 100$. During classification, each feature type is classified by its one-vs-rest SVM to yield SVM scores for a test video. All SVM predictions are subsequently combined via averaging in a late fusion to yield an overall classification of the video according to the maximum score. Note that more advanced late fusion strategies than averaging can be applied to improve performance further, *e.g.* by determining fusion weights via cross-validation on the training data. Furthermore, fusion could also be carried out on the descriptor level or on the encodings level which generally achieves slightly higher performance [4] at the cost of processing very high dimensional vectors for classification.

5. Empirical evaluation

5.1. Datasets and experimental protocols

We evaluate our approach on three widely considered action recognition datasets. The first is HMDB51 [24], which arguably is the most challenging action recognition dataset to date. HMDB51 contains 6766 videos that have been annotated for 51 actions. The authors of the dataset also provide stabilized versions of the videos; however, we only use the original videos in our experiments.

The second dataset we consider is J-HMDB [19], which is a subset of HMDB51 consisting of 21 categories involving only a single person in action. Nevertheless, this dataset remains very challenging, since the excluded categories from the original HMDB51 are mostly facial expressions (*e.g.* smiling), interactions between people (*e.g.* shaking hands) and actions that can only be done in a specific way and thus allow little intra-category variation (*e.g.* a cartwheel).

The third dataset we evaluated our approach on is UCF101 [38], which consists of 13320 realistic action videos in 101 categories and thereby is the largest action recognition dataset to date. It provides huge diversity in terms of action classes, large variations in background, illumination, camera motion and viewpoint, as well as object appearance, scale and pose. However, in terms of scale variations it is less challenging than HMDB due to a fixed resolution of 320×240 for all clips.

For all three datasets, we use the same evaluation protocol as suggested by the authors, which consists of three provided splits into training and test data. Performance is reported as the mean average accuracy over the three splits. We use exactly the same implementation parameters for evaluation on all datasets, as documented in Sec. 4.

5.2. Comparison to ground truth (GT) annotations

Weighting	Features	Descriptors	Acc.
GT-mask[19]	DT	MBH+HOG+HOF+Traj	60.4
GT-mask[19]	GT-flow	MBH+HOG+HOF+Traj	62.4
$\mathcal{S}_\mathrm{T}$	DT	MBH+HOG+HOF	63.3
$\mathcal{S}_\mathrm{T}$	DT	MBH+HOG+HOF+LATE	63.8
$\mathcal{S}_\mathrm{T}$	IDT	MBH+HOG+HOF	64.1
$\mathcal{S}_\mathrm{T}$	IDT	MBH+HOG+HOF+LATE	65.9

Table 1. Mean classification accuracy over three train/test splits on J-HMDB. GT-mask means ground truth foreground puppet mask and GT-flow means ground truth puppet flow annotations.

We begin by comparing our algorithm to an alternative that makes use of groundtruth to restrict action recognition to operate only in foreground regions where actions occur. In this regard, Table 1 reports results for our approach on the elaborately annotated J-HMDB dataset [19]. We investigate the effect of having ideal foreground separation of the actor (GT-mask), as well as ideal trajectories based on ground truth optical flow (GT-flow). The annotations are provided in the form of a puppet mask as *e.g.* shown in Figure 1(d)) and the puppet flow as *e.g.* shown in Figure 1(e); details on puppet annotations are available elsewhere [55]. It is seen that our algorithmically derived saliency weight result at 64.1% mean accuracy actually surpasses that of groundtruth foreground masks, 60.4%.

Interestingly, even giving all dense trajectory descriptors (*i.e.* MBH+HOG+HOF+Traj) the advantage of being computed from groundtruth primitives (*i.e.* GT-flow) only increases their performance to 62.4% mean accuracy. This slight increase by 2% suggests that all the trajectory descriptors share similar information and therefore are not very complementary [19]. In contrast, augmenting the IDT descriptors with our novel LATE descriptor, (19), further improves performance of our approach to 65.9%.

5.3. Comparison to alternative saliency approaches

Weighting	Features	Descriptors	Accuracy
[39]	STIP	HOG+HOF	69.30
$\mathcal{S}_\mathrm{T}$	STIP	HOG+HOF	77.33
$\mathcal{S}_\mathrm{T}$	Grid (Sec. 3)	LATE	78.22
$\mathcal{S}_\mathrm{T}$	IDT	MBH	89.78
$\mathcal{S}_\mathrm{T}$	IDT	HOG	88.89
$\mathcal{S}_\mathrm{T}$	IDT	HOF	84.89
$\mathcal{S}_\mathrm{T}$	IDT	MBH+HOG+HOF	90.89
$\mathcal{S}_\mathrm{T}$	IDT + Grid (Sec. 3)	MBH+HOG+HOF+LATE	92.22

Table 2. Mean classification accuracy for weighted encoding of different features on a subset of HMDB51 [39], only including the classes Biking, Golf swing, Pull ups, Horse riding and Basketball.

We now explicitly evaluate our spacetime saliency measure, $\mathcal{S}_\mathrm{T}$, in comparison to two alternative approaches that previously have employed algorithmically derived saliency for action recognition.

The first alternative [39] is based on a graph-based notion of saliency defined over colour and optical flow gradients, as noted in Sec. 1. The alternative approach was only evaluated against a subset of HMDB restricted to the action classes of Biking, Golf swing, Pull ups, Horse riding and Basketball. Correpsondingly, our evaluation is restricted in the same fashion in this comparison (note that this subset is idiosyncratic, not that defined by J-HMDB). Results are shown in Table 2. Results for the alternative were only reported for features sampled according to STIP [25] with HOG [11] and HOF [25] descriptors. A wider variety of feature sampling and descriptor approaches are presented for the proposed approach. In all cases, the proposed spatiotemporal saliency-based approach outperforms the alternative, with improvements ranging between approximately 9 and 23% in mean accuracy. Moreover, it is noteworthy that LATE provides a more efficient representation of video, *e.g.*, for the entire HMDB51, the LATE descriptors consume an order of magnitude less memory than the IDT representation.

A second particularly interesting point of comparison

is to the other approach that makes use of spatiotemporal saliency to weight feature pooling [3]. In that case, three types of saliency measures are used to pool LLC [45] encoded DT [42] features in a spacetime pyramid and perform classification with a weighted SVM model, as discussed in Sec 1. In Table 4 one observes that this alternative method (DT + Saliency'13 [3]) is outperformed by the proposed approach by 51.8 vs. 57.4 mean accuracy on the entire HMDB51.

5.4. Evaluation of Feature Aggregation Schemes

Method	Features	Aggregation	Descriptor		
			HOG	HOF	MBH
[42]	DT	FV	33.3	36.9	44.6
[51]	DT	SDV	33.1	37.3	44.3
[51]	DT	STP [51]	34.4	38.1	46.9
[51]	DT	SSCV	36.9	39.7	48.0
[43]	IDT	FV	40.2	48.9	52.1
Proposed	DT	$\mathcal{S}_\mathrm{T}$	43.9	44.5	53.2
Proposed	IDT	$\mathcal{S}_\mathrm{T}$	45.1	51.9	54.6

Table 3. Mean classification accuracy for different aggregation schemes of various features and descriptors on HMDB51.

Recently, Yang and Tian [51] compared the performance of different aggregation schemes for features sampled according to dense trajectories (DT) [42] and Improved Dense Trajectories (IDT) [43] with various descriptors on the HMDB51 dataset. In Table 3, we show the performance of our weighted FV aggregation based on motion saliency for the same descriptors. Our approach is compared to those originally presented for DT [42] and IDT [43], which also used Fisher vector encoding (FV), but without any notion of saliency weighting. We also show the performance of additional novel feature aggregation approaches for action recognition [51]. Interestingly, it is seen that merely changing the feature sampling from DT to IDT allows FV to outperform the other aggregation schemes operating over DT. Further improvement is had by including our saliency weighted FVs, which yields the best overall performance.

5.5. Comparison with the state-of-the-art

In Table 4 we compare our approach against the state-of-the-art in action recognition. For both the HMDB51 and UCF101 datasets, direct comparison between our results for IDT + $\mathcal{S}_\mathrm{T}$ (penultimate row) with those for the alternatives documents the performance change provided by inclusion of the proposed saliency weighting. It is seen that improvement is provided on HMDB51, while results are essentially the same as the best alternatives for UCF101 as performance is becoming largely saturated on that dataset. Interestingly, it also is seen that additional inclusion of the LATE descriptors further boosts performance on all datasets (bottom row), which suggests a degree of complementarity between the novel LATE descriptors and the more standard HOG, HOF and MBH descriptors. Finally, it is notable that by combing Fisher vectors (FV) with Stacked Fisher Vectors (SFV) even further improvement can be had with IDT to yield mean accuracy of 66.8% on HMDB [30], albeit at the cost of doubling the representation dimensionality and large computational complexity. Since our approach consistently improves IDT with FV, we anticipate that it would even further improve the FV/SFV combination.

HMDB51		UCF101	
DT + Saliency'13 [3]	51.8		
DT + FV'13 [43]	52.2		
DT + SSCV'14 [51]	53.9		
DT + Actons'13 [53]	54.0	ST ConvNet'14 [20]	65.4
DT + MVSV'14 [4]	55.9	DT + MVSV'14 [4]	83.5
IDT + SFV'14 [30]	56.2	DT + bimodal'14 [48]	84.2
IDT + FV'13 [43]	57.2	IDT + FV'13 [44]	85.9
IDT + DaMN'14 [16]	57.9	IDT + DaMN'14 [16]	87.0
TS ConvNet'14 [37]	57.9	TS ConvNet'14 [37]	87.6
DT + $\mathcal{S}_\mathrm{T}$	57.4	DT + $\mathcal{S}_\mathrm{T}$	84.3
DT + LATE + $\mathcal{S}_\mathrm{T}$	58.5	DT + LATE + $\mathcal{S}_\mathrm{T}$	85.5
IDT + $\mathcal{S}_\mathrm{T}$	61.5	IDT + $\mathcal{S}_\mathrm{T}$	86.4
IDT + LATE + $\mathcal{S}_\mathrm{T}$	62.2	IDT + LATE + $\mathcal{S}_\mathrm{T}$	87.7

Table 4. Mean classification accuracy over three train/test splits on HMDB51 and UCF101. The bottom four rows show results for the proposed approach using DT and IDT feature sampling with HOG + HOF + MBH descriptors (DT + $\mathcal{S}_\mathrm{T}$ and IDT + $\mathcal{S}_\mathrm{T}$) as well as with further addition of LATE descriptors (DT + LATE + $\mathcal{S}_\mathrm{T}$ and IDT + LATE + $\mathcal{S}_\mathrm{T}$). Results for alternative approaches are shown in the upper portion of the table.

6. Conclusion

This paper has presented a novel approach to pooling within the BoW framework applied to action recognition. Pooling is performed with a novel saliency-based weighting function that has highest values in regions of foreground motion where an action is most likely to occur. The approach can be applied in conjunction with any locally defined features and encoding methods. Here, it has been instantiated using Improved Dense Trajectories (IDT) and novel Locally Aggregated Temporal Energy (LATE) features. LATE involves neighborhood aggregation of local temporal energy to capture the spatiotemporal layout of the individual measurements, *e.g.* to capture the relative arrangement of different action parts. During encoding we enhance the contribution of local features by weighting with their respective saliency.

The overall action recognition system is competitive with and can even improve over the previous state-of-the-art on the HMDB51, J-HMDB and UCF101 datasets. These results suggest the importance of explicitly concentrating processing on regions where an action is likely to occur during recognition.

Acknowledgments: This work was supported by the Austrian Science Fund (FWF) under project P27076.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

[2] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, 2012.

[3] N. Ballas, Y. Yang, Z.-Z. Lan, B. Delezoide, F. Preteux, and A. Hauptmann. Space-time robust representation for action recognition. In *Proc. ICCV*, 2013.

[4] Z. Cai, L. Wang, X. Peng, and Y. Qiao. Multi-view super vector for action recognition. In *Proc. CVPR*, 2014.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014.

[6] W. Chen, C. Xiong, R. Xu, and J. Corso. Actionnes ranking with lattice conditional ordinal random fields. In *Proc. CVPR*, 2014.

[7] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *Proc. CVPR*, 2011.

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with Fisher vectors. In *Proc. ICCV*, 2013.

[9] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[10] N. Dalal. Finding people in images and videos. *PhD thesis, Institut National Polytechnique de Grenoble*, 2006.

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

[12] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. ECCV*, 2006.

[13] K. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *PAMI*, 35(3):527–540, 2013.

[14] A. Gaidon, Z. Harchaoui, and C. Schmid. Activity representation with motion hierarchies. *IJCV*, 107(3):219–238, 2014.

[15] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *PAMI*, 34(10):1915–1926, 2012.

[16] R. Hou, A. R. Zamir, R. Sukthankar, and M. Shah. Damn discriminative and mutually nearest: Exploiting pairwise category proximity for video action recognition. In *Proc. ECCV*, 2014.

[17] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, 2012.

[18] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, 34(9):1704–1716, 2012.

[19] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proc. ICCV*, 2013.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. CVPR*, 2014.

[21] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proc. BMVC*, 2008.

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

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[24] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proc. ICCV*, 2011.

[25] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. CVPR*, 2008.

[26] I. Laptev, M. Piccardi, M. Shah, R. Sukthankar, Y. Jiang, J. Liu, and A. Zamir. THUMOS: ICCV 2013 workshop on action recognition with a large number of classes. 2013.

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

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

[29] A. V. Oppenheim, R. W. Schafer, J. R. Buck, et al. *Discrete-time signal processing*, volume 5. Prentice Hall, Upper Saddle River, New Jersey, USA, 1999.

[30] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked Fisher vectors. In *Proc. ECCV*, 2014.

[31] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proc. CVPR*, 2012.

[32] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc. CVPR*, 2007.

[33] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

[34] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 34(3):601–614, 2012.

[35] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *Proc. BMVC*, 2013.

[36] E. Shechtman and M. Irani. Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? *PAMI*, 29(11):2045–2056, 2007.

[37] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. NIPS*, 2014.

[38] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions calsses from videos in the wild. Technical Report CRCV-TR-12-01, UCF Center for Research in Computer Vision, 2012.

[39] W. Sultani and I. Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *Proc. CVPR*, 2014.

[40] M. Ullah, S. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *Proc. BMVC*, 2010.

[41] E. Vig, M. Door, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye-movements. In *Proc. ECCV*, 2012.

[42] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recog-

nition. *IJCV*, pages 1–20, 2013.
[43] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. ICCV*, 2013.
[44] H. Wang and C. Schmid. LEAR-INRIA submission for the THUMOS workshop. In *ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013.
[45] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Proc. CVPR*, 2010.
[46] L. Wang, Y. Qiao, and Z. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *Proc. CVPR*, 2013.
[47] B. Watson and A. Ahumada. A look at motion in the frequency domain. In *Proceedings of the Motion Workshop*, 1983.
[48] J. Wu, Y. Zhang, and W. Lin. Towards good practices for action video encoding. In *Proc. CVPR*, 2014.
[49] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proc. CVPR*, 2013.
[50] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. CVPR*, 2009.
[51] X. Yang and Y. Tian. Action recognition using super sparse coding vector with spatio-temporal awareness. In *Proc. ECCV*, 2014.
[52] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *Proc. ECCV*, 2010.
[53] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu. Action recognition with actons. In *Proc. ICCV*, 2013.
[54] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *Proc. CVPR*, 2014.
[55] S. Zuffi and M. J. Black. Puppet flow. Technical Report TRIS-MPI-007, MPI for Intelligent Systems, 2013.