Evaluating Descriptors for Human Action Recognition in Hollywood Videos: Static Scene Representations vs. Local Space-Time Features

Michael C. Hughes
Brown University
115 Waterman St
Providence, RI 02906
mhughes@cs.brown.edu

Abstract

Local space-time descriptors have become a successful video representation for human action classification. Recent investigations suggest that for diverse, realistic datasets, computing these descriptors on a dense, regular grid outperforms sparse interest point methods. We suggest this dense sampling succeeds by capturing meaningful global context in addition to local dynamics. In this investigation, we evaluate this claim by comparing state-of-the-art space-time descriptors (HOG and HOF) with static scene descriptors (GIST and dense SIFT) in a bag-of-features classification task on the Hollywood2 actions dataset. Results indicate that for one-third of tested action categories, static scene descriptors can outperform dynamic ones. We also show that combining static and dynamic descriptors yields even further improvements, suggesting novel avenues for further research in video representation.

1. Introduction

Recognizing natural human actions under realistic video settings remains a foremost challenge of computer vision. The essential task under study here is to assign a predefined categorical label to a query video segment. For example, given a short clip of Forrest Gump sprinting across a field, we wish the system to identify it as an instance of the Run category.

Many sophisticated automated representations are possible for this task, such as object recognition, tracking, etc. However, one line of recent research has demonstrated the success of local space-time descriptors, which produce feature vectors extracted from raw pixel information with minimal preprocessing. This approach provides a representation that is relatively independent of shifts and scales and is less vulnerable to background clutter, occlusions, and other motions compared to object-level representations.

These space-time descriptor representations produce successful results in human action classification in a variety of contrived datasets such as KTH which show single actors exhibiting staged, repetitive exercises like running or boxing. More recently, the release of diverse datasets of thousands of videos have provided new benchmarks for analyzing these representations. A prominent dataset is Hollywood2, a collection of video segments extracted from major motion pictures [5]. For a quick visual impression, see Figure 1 or view the online preview video[1]. This dataset presents a host of challenges: far more categories than other datasets, significant intra-class variations, occlusions, camera motions, and diverse poses and backgrounds. Beneficially, however, it has suggested a strong qualitative shift in the approach used for classification.

Traditionally, work on video action classification focused a great deal on *interest point* detection. Investigators attempted to find sparse points in spatio-temporal domain that could provide strong discriminative clues for classification, and then deployed descriptors only on the selected keypoints. However, a confluence of recent work suggests that instead of *sparse* sampling, it may be more beneficial to adopt a *dense* approach, at least on realistic videos. Three lines of evidence support this conclusion.

First, we note the release the first large-scale cross-dataset evaluation of different interest point detectors and descriptors [7]. Wang *et al*. evaluate four different feature detectors combined with six descriptors on three distinct datasets. Their results demonstrate that that "regular sampling of space-time features consistently outperforms all tested space-time interest point detectors for human actions in realistic settings." This is likely because dense sampling captures both complete descriptions of the actions as well as rich context information.

Second, a recent paper by Marszalek *et al*. showcases the inherent correlations between human actions and scene

[1] http://www.youtube.com/watch?v=1g5G5Dmvt3o

context information [5]. From an initial qualitative observation that "eating often occurs in a kitchen, but running is more common outdoors," the authors move to develop a joint framework for action and scene recognition for the Hollywood2 dataset. Explicitly modelling both scene type and action type, the authors show that contextual relations improve both tasks.

Third, stepping back from video analysis we notice a broader trend in the computer vision literature that favors dense sampling. Dense SIFT features have been particularly beneficial in large-scale scene classification and object recognition tasks [4]. It seems that as computer vision moves from hand-crafted sterile datasets to rich natural images and videos from the wild, dense representations are necessary to boost performance.

Motivated by this shift toward dense sampling, we investigate global context for action recognition one step farther. Instead of just capturing local space-time descriptions at a dense grid of points, we ask: how well can static, scene-wide descriptors perform at action recognition in challenging settings?

We study two popular representations for global scene matching in static images: the GIST descriptor and dense SIFT descriptor [6, 4]. To apply these methods to videos, we introduce a novel video description approach in which a video is turned into a short sequence of static frames sampled over wide delays in time, and each of these frames is then fed to a static image descriptor. The outcome is a bag-of-features model for the scenes sampled from a video.

We compare the global features to two prominent space-time descriptors: Histogram of Gradients (HOG) and Histogram of Optical Flow (HOF), both originally proposed in [3]. We choose these because they were the top-ranked descriptors in the large-scale evaluation of Wang *et al*. Other descriptors are possible, for example the Cuboids descriptor [2]. However, most operate on gradients of the spatio-temporal volume and thus does not seem to capture qualitatively different information than the HOG descriptor. This fact together with the top-rated performance of HOG and HOF descriptors in multi-dataset evaluations ([7]) suggests that HOG and HOF alone are good choices for this task.

The remainder of this report is organized as follows. Section 2 introduces the dataset under investigation, Hollywood2, and describes preprocessing done on this dataset. Section 3 addresses the descriptors used to represent each video in the dataset, providing implementation details for how global scene structure and local space-time motion were represented. Section 4 describes the bag-of-features classification approach in detail. Section 5 shares results of classification task, and Section 6 discusses lessons learned.

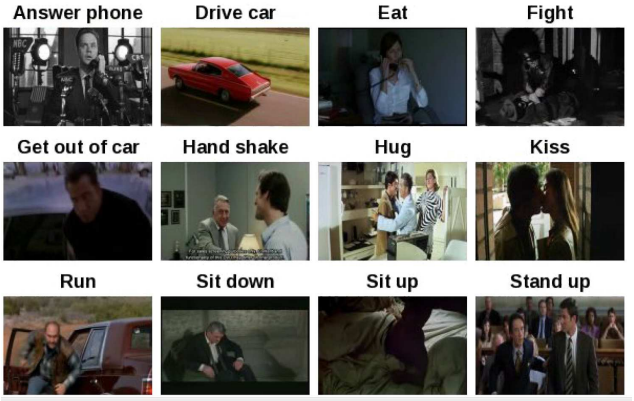


Figure 1. Hollywood2 dataset: 12 action categories with diverse visual characteristics.

2. Data

The Hollywood2 Actions dataset[2] consists of 12 action categories in 1707 short clips (each roughly 3-30 seconds long) cut from 69 different wide-release Hollywood motion pictures, ranging from "Bruce Almighty" to "Forrest Gump" to "Pirates of the Carribbean." The data is split into a Training set (823 videos) and a Test set (844 videos). Each set draws from a distinct set of movies, so instance-level correspondence between any Training and Test videos is highly unlikely. Two versions of the dataset are available, and we choose the manually verified one instead of the automatically generated one (see **??**).

2.1. Down-sampled Preprocessing

To make computation affordable and timely, we chose to down-sample the raw videos from their original encoding. In original form, each video had minimum resolution 500x300 and 25 frames-per-second. We chose to capture every other frame in grayscale at half spatial resolution for use in our experiments. We also capture a maximum of 400 frames per video (less than 10% of videos exceed this). All descriptors studied here thus operated on grayscale video with minimum resolution 250x150 and 12 fps.

Although we expect this will reduce overall classification performance, we assert that the focus of this investigation is a relative comparison of different descriptor classes rather than an attempt to achieve state-of-the-art results. We have no reason to believe any descriptor is disadvantaged by this scheme, so we favor it for its significant computational efficiency.

3. Descriptors

Here we describe both the nature of our chosen representations and implementation details. Specifically novel is the

[2] http://www.irisa.fr/vista/actions/hollywood2/

method employed to capture static scene-level information for a video.

3.1. Global Scene descriptors

Many choices for global scene representations exist. We chose GIST and dense SIFT for their acclaim and qualitative differences. For implementation of both descriptors, we used the code provided in the LabelMe MATLAB toolbox[3].

3.1.1 GIST

Oliva and Torralba first proposed the GIST descriptor as a compact representation inherent characteristics of a scene such as ruggedness and openness [6]. To compute the GIST for a video, we preprocess each input frame by center cropping and resizing to produce a 128-by-128 pixel square image. The image is then subdivided into 4 cells along each spatial dimension and for each one Gabor filter responses are computed at 8 orientations per scale (these are default parameters for LabelMe implementation). The outcome is a 512-dimensional GIST feature vector for each sampled frame. We thus capture only one "visual word" per frame in this case.

3.1.2 SIFT

While SIFT was originally well-known as a descriptor applied to sparse interest points, later investigations found additional success applied the descriptor to densely sampled patches across an entire image frame [4]. In our implementation, we compute a SIFT feature vector for each patch in a dense grid of square pixel patches each of length 20 pixels with 50% overlap. Each patch produces a 128-dimensional feature vector, and we count each patch as one "visual word" for the bag-of-words representation of the video. Although previous work suggests sampling at multiple scales might boost performance, we did not consider this extension due to time constraints.

3.1.3 Sampling Static Images from Video

Both GIST and dense SIFT descriptors require static images as input, not video. We simply extract frames periodically from the video at a constant sampling rate, as shown in Figur e 2. We sample every 10th frame of a video, corresponding to a frame rate of roughly 1 per second due to our pre-processing (Section 2.1). We hope this rate provides a balance between missing crucial scene changes due to camera panning or zooming, and capturing too much redudant information in static camera scenes. Perhaps alternative approaches that take into account shot boundaries, camera motion, and other relevant factors may be appropriate for future investigations, but we do not consider them here.

[3]http://labelme.csail.mit.edu/LabelMeToolbox/

Figure 2. Example frame sampling for global static descriptors.

Each sampled frame is fed to the relevant descriptor, and the resulting feature vector(s) are concatenated across all frames to build a bag-of-features representation for that video.

3.2. Dynamic Spatio-temporal descriptors

As input, local space-time descriptors consume the frame-by-frame raw pixel intensities for an entire input video, and produce a compact histogram representation of each space-time volume in the video. We densely sampled space-time volumes from each video with spatial dimension $S = 20$ pixels and temporal dimension $T = 10$ frames along a grid with 50% overlap. These parameters are roughly consistent with implementations of [7].

Note that most previous investigations compute a feature histogram for spatially and temporally subdivided cells within each sampled space-time volume, and also compute results for multiple scale volumes centered at each sample point [7]. Due to time constraints and computational expense, most results reported here do perform these modifications and instead produce just one feature histogram per volume. See Section 5.4 for evaluation of this design choice.

3.2.1 HOG: Histogram of (Spatial) Gradients

For a particular space-time volume, the HOG describes how local spatial gradients change orientation over time. For a given space-time volume, the gradient orientation at each pixel is binned into one of 4 orientations [7]. These correspond to vertical edges, diagonal forward slash edges, horizontal edges, and diagonal back-slash edges. This is visuallized in Figure 3. A histogram for the entire volume is computed by aggregating the counts in each bin over all pixels in the space-time volume. We finally normalize the resulting histogram by Euclidean distance.

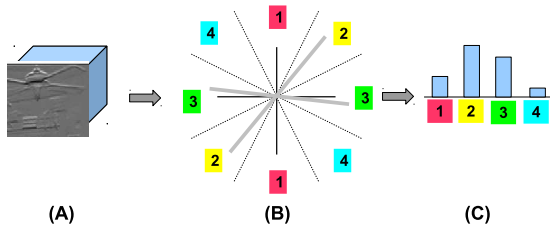


Figure 3. Histogram of Gradients (HOG) descriptor.

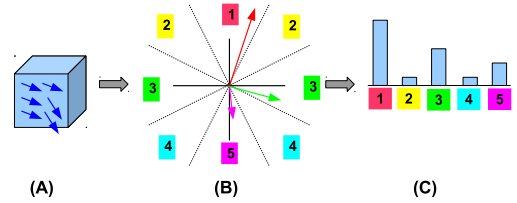


Figure 4. Histogram of Optical Flow (HOF) descriptor.

3.2.2 HOF: Histogram of Optical Flow

The HOF descriptor proposed by [3] indicates how observed motions in a local space-time volume change over time. Given an input video, we first compute the optical flow between adjacent frames $i, i+1$ using a fast, parallel implementation of the Lucas & Kanade algorithm [4]. For each space-time volume, we compute a 5 bin histogram (like [7]) where each bin corresponds to a horizontally symmetric velocity orientation, as originally proposed in [1]. Intuition suggests that this symmetry matches the way we classify our visual world, as we do not distinguish walking to the left and walking to the right as distinct actions but we do consider standing up versus sitting down to be distinct. As figure 4 indicates, we aggregate histogram counts by choosing the bin for each pixel's flow vector and adding the magnitude of that flow vector. We finally normalize the resulting histogram using Euclidean distance.

4. Bag-of-Features and Classification

For each individual video in the dataset, we obtain a set of feature vectors as output from each distinct descriptor. Note that each video has on the order of 10^1 GIST features (one per sampled frame), but 10^3 SIFT features (one for each dense patch at each sampled frame) and 10^4 HOG and HOF features (one for each densely sampled space-time volume). To perform classification, we first compute a single bag-of-features histogram for each video, and then use this compact representation as a basis for classification. We explore both nearest neighbors methods and an SVM classifier.

[4] http://vision.ucsd.edu/ pdollar/toolbox/doc/images/optFlowLk.html

4.1. Bag-of-Features

We assemble a visual vocabulary for each feature by first obtaining a sample of 100000 features uniformly at random from all training videos (since GIST has far fewer features per video we only sample 5000). These are then clustered via k-means given a fixed vocabulary size V and distance metric. We found $V = 500$ and L1 distance gave a good balance between computational efficiency and classification performance. Memory limitations prevented testing $V = 4000$ as used in [7].

We use the V cluster centroids output by k-means as the visual words. For each video, we assign each feature vector to its nearest visual word via the same distance metric used in vocabulary construction. We then aggregate counts of word occurance into a V-bin histogram. Each occurance histogram $H_i = \{h_{in}\}$ is normalized so the sum of bin counts is unity: $H_i = h_{in} / \sum_n h_{in}$.

4.2. Nearest Neighbors Classifier

As a classification baseline, we first adopted a simple nearest neighbors approach to classifying a query video. Given a video from a test set to label, we compare its bag-of-features histogram to those of 823 Training videos and extract the N nearest neighbors using a specified distance metric. We found L1 distance to yield good results. We hope in future work to compare other metrics (such as χ^2 distance) more thoroughly. In all reported NN results, we use $N = 50$, which yielded good quality results for all descriptors.

4.3. Non-linear SVM Classifier

We also employ a non-linear SVM with a χ^2 kernel, which is the common choice in human action recognition using bag-of-features pipeline [3, 7]. The kernel matrix K is defined by

$$K(H_i, H_j) = \exp\left(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in}-h_{jn})^2}{h_{in}+h_{jn}}\right) \quad (1)$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are frequency histograms of word occurances, V is vocabulary size, and A is mean value of the pairwise distances between all training examples.

For each action class, we train a separate binary SVM classifier in a one-against-rest approach.

5. Results

Following previous action classification research, we use average precision (*AP*) as the metric for comparing performance on a particular action class. A perfect classifier has $AP = 1$, while a random classifier should expect AP

roughly equal to the fraction of all examples that are true positives. When studying overall performance of a particular feature or classifier, we report only the mean average precision across all categories.

5.1. NN vs. SVM

As a sanity check and baseline, we compare nearest neighbor classification performance to the χ^2 SVM using the metric of mean average performance across all action classes. We see in Figure 5.1 that the SVM results are significantly better than NN results across all descriptors. We thus use the SVM classifier for the remainder of results reported here.

5.2. Global vs. Local Descriptors

Our first experiments examine how well static scene descriptors (GIST and SIFT) fare against local space-time descriptors in action classification. Figure 6 displays average precision results for each of the 12 action classes present in the Hollywood2 dataset using the SVM classifier. We draw several qualitative lessons from this plot. First, static scene descriptors *outperform* dynamic counterparts on four of the twelve classes. GIST fares best at the AnswerPhone category, while SIFT is best at GetOutCar, HandShake, and Run. These classes seem to be highly correlated with particular scene types (phone conversations occur in an office) or perhaps particular objects (e.g. GetOutCar videos always feature cars). We suggest this evidence lends credibility to our hypothesis: global scene representation can be significantly important for action classification.

Note however that local dynamic descriptors are best a majority of the time. HOF descriptor seems especially powerful in actions that occur in a wide range of scenes but have strong motion patterns (e.g. StandUp and SitDown). Also, some classes that we might naively expect scene representation to do well at (e.g. Eat) actually show the best results with dynamic descriptors.

5.3. Combining Global and Local Descriptors

Given the success of global scene features at classification, we conducted further experiments to determine if a *combination* of global and dynamic descriptors might boost performance beyond solitary capabilities. For these experiments, to represent a single video using both features A and B, we simply concatenate the bag-of-features histograms H_A and H_B together into one histogram with $2V$ bins, the first half from A and the second from B. We perform no additional normalization.

We compute the mean average precision across all 12 action classes for each feature combination. In figure 7, we compare individual feature performance against possible pair-wise combinations of features. We notice that in general, adding dense SIFT to any other feature appears to boost performance significantly. The combination of SIFT+HOF is actually slightly better than HOG+HOF, the combination that performed best in Wang *et al.*'s comparative study. These results further support the conclusion that static global scene representations (namely SIFT) can provide meaningful information for classifying human actions. Adding GIST doesn't seem to improve any other feature, indicating that it either might be less relevant, or that the dataset is not large enough for reliable scene matching.

Descriptor	Mean AP	Mean AP Multiple Scales	Mean AP with Subdivisions
HOG	0.26	0.24	**0.37**
HOF	0.32	0.32	**0.34**

Table 1. Comparing our descriptor's classification performance (2nd column) to enhancements suggested in [3] (col 3,4)

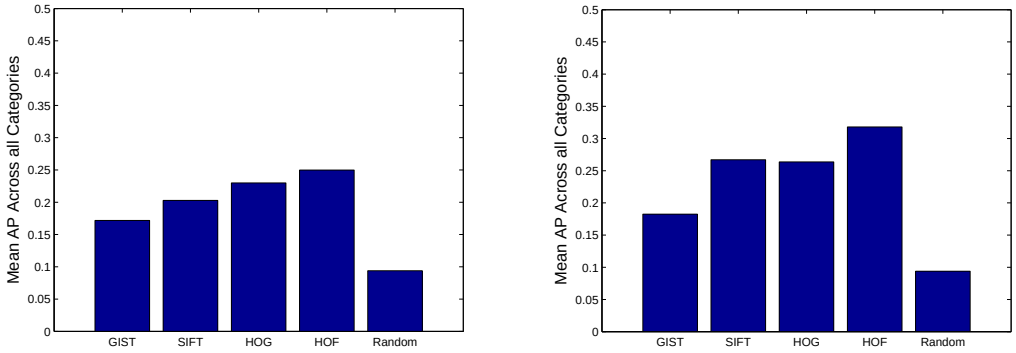
5.4. Extending local space-time descriptors

Note carefully that the dynamic descriptors used here are not exactly those for which results are reported in [7] or related papers. Particularly, we did not consider the multi-scale or space-time volume subdivisions that previous authors have, in order to simplify analysis. This means our absolute AP results are not necessarily comparable to theirs.

We have run a brief test for comparison to help gauge how much their modifications improved classifier performance. Results are displayed in Table 1. Particularly we find, that computing feature vectors at multiple scales for each dense point in the sampling grid does not seem to significantly increase performance of either HOG or HOF descriptors. Alternatively, we consider using only one scale (the same used throughout the paper) but dividing each volume into smaller cells. We use 2 in each spatial dimension and 2 in time, resulting in 8 unique feature vectors (one per cell) computed for each volume. We concatenate these vectors together to produce a single 32-dimensional vector for the volume's HOG and 40-dimensional vector for the volume's HOF. A glance at the table shows significant performance gains. The HOG descriptor increases from 0.26 to 0.37 using volume subdivisions! Future analysis should probably consider volume subdivisions, and perhaps a combination of volume subdivisions and multiple scales.

6. Conclusion

The evidence accumulated in this investigation suggests that global scene representation for video can contribute meaningfully to overall classifier performance and actually be more discriminative for specific action classes that have strong correlations with scene statistics. These results are remarkable because in our implementation dense SIFT captures roughly 1/10 of the feature vectors com-



(a) 50 NN (b) SVM

Figure 5. Classifier comparison: NN vs. SVM.

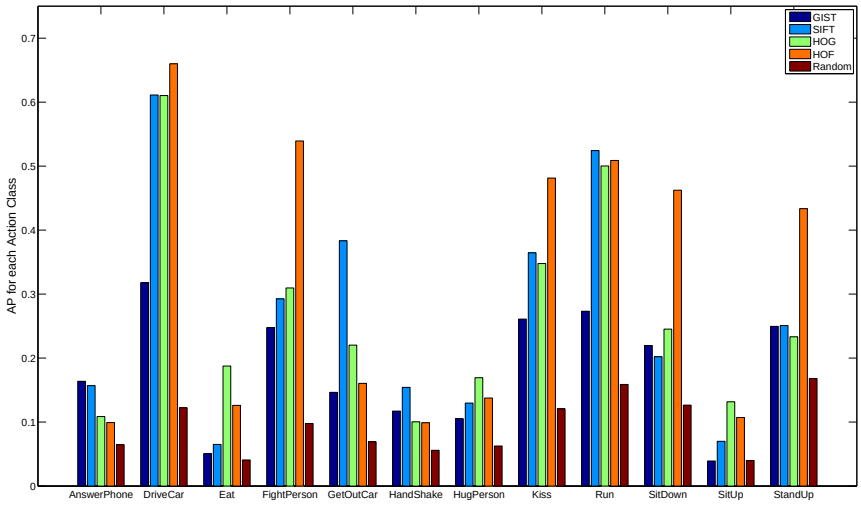


Figure 6. Average Precision results for each Hollywood2 Action Class using SVM classifier.

pared to HOG/HOF descriptors, while GIST captures fewer than 1/100. We especially recognize the exceptional performance of dense SIFT features as a noteable outcome of this investigation.

6.0.1 Further Investigations

Perhaps most interesting is the further study of novel descriptors that combine global scene statistics and local dynamics. Subdividing space-time volumes also appears to be a promising direction. It is unclear if a simple concatenation of separate global and local bag-of-feature histograms is the best solution for this task, or if alternative methods might achieve better results.

The influence of dataset size on performance is also notable for study. Particularly, we expect scene matching to get much better given 1000s or even millions of sample videos rather than the 800 trained here. A complementary investigation could study what happens as the number of action classes grows. We expect that as scenes are no longer associated with only one or two actions, the performance of scene representation will decay rapidly compared to dynamic descriptors. This fact should force investigators to pay close attention to relationship between the intended classification task, the data available, and the representation to employ.

In closing, we suggest that densely-sampled local features combined with some static global representation can offer significant benefits to action classification in Hollywood movies videos. As computer vision investigators continue to consider realistic datasets, it seems that representa-

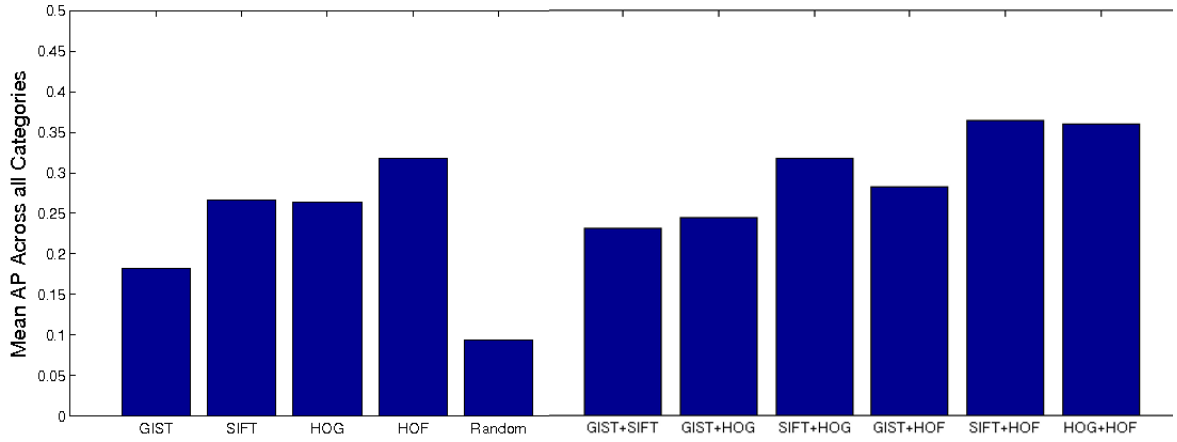


Figure 7. Individual descriptor performance compared to pairwise combinations.

tions must capture all available context in order to succeed.

References

[1] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009. 4

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005. 2

[3] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2008. 2, 4, 5

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing scene categories. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2005. 2, 3

[5] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009. 1, 2

[6] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vision*, 42:145–175, 2001. 2, 3

[7] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, page 127, sep 2009. 1, 2, 3, 4, 5