Activity Recognition in Videos using 3-D Salient Point Signatures

Rohan Jain
Vision Research Lab
University of California, Santa Barbara
rohanjain@umail.ucsb.edu

Abhejit Rajagopal
Scientific Computing Group
University of California, Santa Barbara
abhejit@ece.ucsb.edu

Abstract

In this paper, we review state of the art methods for human activity recognition using moderate frame-rate videos. In particular, we implement a modular activity recognition system that provides a suitable testbed for evaluating the performance of various detectors, descriptors, and classifiers. For our work, we have used the publicly available KTH dataset and evaluated our algorithm on a multi-class classification problem. In our experiments, we have focused on computer vision techniques for improving the performance of activity recognition rather than the specific classification tool employed. Our optimal method detects spatio-temporal interest points in a subject video using SURF, and characterizes these keypoints using local HOG3D features to produce a Bag of Visual Words (BOVW) representation for each activity. This histogram-like feature is then combined with a global R-transform feature via an innovative feature fusion technique and finally incorporated into a supervised learning framework. Our implementation achieves an average accuracy of ***84.1%*** *on the KTH Dataset, while state of the art methods report an average accuracy of 92.1%.*

1. Introduction

Activity recognition is a compelling field within computer vision because it provides the ability to quickly process, annotate, and organize video information automatically. Videos in this context refer to a series of pictures (frames), typically in the optical regime, that are sampled sufficiently fast as to provide both spatial and temporal information to the viewer. Unlike 1-dimensional signals, which may contain important amplitude or frequency information, 2-dimensional pictorial signals are typically porous (sparse) in the information content required for activity recognition. Rather than exhibiting large information redundancies, such as those explained by an optimal set of basis functions, activities can be typically recognized by tracking only a few points of interest [5]. Prime examples of this methodology include tracking the back-and-forth motion of a boxer's hands, or the up-and-down motion of a jogger's legs. Obviously, fitting these activities to such a simple 1-dimensional model will result in considerable degeneracy; even still, in such a case context-specific information can be used to differentiate between boxing and activities with similar motion, such as sawing or rowing.

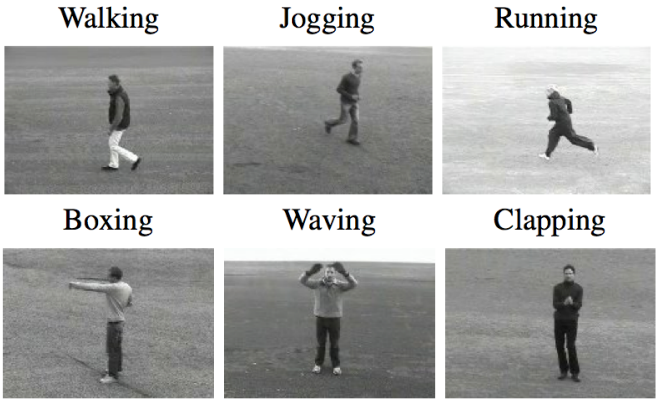


Figure 1. Six acitivies are included in the KTH dataset [7][10].

In this paper, we first review state of the art methods of activity recognition incorporating spatio-temporal volumes. A majority of these techniques employ a salient point detector, a local or global feature descriptor, and a multi-class classifier. Next, we provide an overview of our own implementation, inspired by that of [10]. Here, we select two video descriptors, the R-transform (a rotation and scale invariant derivative of the Radon transform) [10] and HOG-3D (spatio-temporal histogram of oriented gradients)[4], as 3-D salient point signatures. Finally, using our modular implementation, we present the results of various detector, descriptor, and classifier schemes.

2. Related Works

The problem of activity recognition in videos has been well addressed within the computer vision community in the recent past. A variety of spatio-temporal interest point

detectors have been proposed. The three-dimensional Harris detector proposed by Laptev and Lindeberg [5] extends the functionality of the popular Harris corner detector to three dimensions. Dollar et al. [3] presented a cuboid detector and descriptor pair which uses temporal Gabor filters to identify keypoints in a video and PCA-SIFT in the local 3D neighborhood to characterize them. The Hessian detector was proposed by Willems et al. [9] as a spatio-temporal extension of the Hessian saliency measure used in [6] for blob detection in images. The detector measures the saliency using the determinant of the 3D Hessian matrix.

The HOG/HOF descriptors introduced by Laptev et al. in [5] characterize local motion and appearance by computing histograms of spatial gradient and optic flow accumulated in space-time neighborhoods of detected interest points. The HOG3D descriptor was proposed by Klaser et al. [4] and is one of the descriptors selected in this work. HOG3D is based on local histograms of 3D gradient orientations and can be seen as an extension of the popular SIFT descriptor [20] to video sequences. Shah et al. [8] extended the SIFT descriptor to 3D using a similar approach. Willems et al. [9] proposed the extended SURF (ESURF) descriptor which extends the image SURF descriptor [2] to videos.

The commonality in all these approaches is a salient-point detector, spatio-temporal descriptor, and a classification tool. We base our approach as elucidated in the next section on the work by Yuan et al. [10], which uses the therein-defined R transform and BOVW features for action recognition by using one feature to get the context and the other feature to calculate a context-aware kernel for carrying out action classification.

3. A Modular Activity Recognition System

Our method calculates a global and a local descriptor of visual information for each video and intelligently combines the information available to characterize and classify the videos according to the activity that they represent. Specifically, the R feature captures the global geometrical distribution information, while the BOVW representation increases the discriminability of actions using local features[11]. It worth noting that whilst designing our own action recognition implementation, we first developed a modular code-base which allows for the substitution of various subroutines for the detector, descriptor, and classifier. Using this we were able to verify the improvement in using said features and fusion schemes (Section 4).

3.1. 3D R Transform

The 3D Discrete Radon Transform has been used to characterize three dimensional models [1]. The 3D R transform is a variation of the radon transform that is more robust to translation, rotation and scale changes [10]. In a digital implementation, these transforms must be discretized first. The 3D discrete Radon transform is defined by summing the interpolated samples of a discrete 3D array lying on planes which satisfy certain constraints. Given $[x_j]_{j=1}^{J}$ be all the points in the model M, the 3D discrete Radon transform of the 3D model $f(x)$ is defined as follows:

$$T_f(\rho,\theta,\phi) = \sum_{j=1}^{J} f(x_j,y_j,t_j)\cdot \delta(x_j cos\phi sin\theta + y_j sin\phi cos\theta + t_j cos\theta - \rho) \quad (1)$$

where $\boldsymbol{x_j} = (x_j, y_j, t_j)$ represents a point in the spatio-temporal volume. This can be easily calculated, but is not invariant to translation, scaling, or rotation. To overcome this, we employ the 3D *R* transform, which is the integral of the square of the 3D Radon transform over the parameter ρ. In a sense, we are removing the shift and scale dependence of the Radon transform by integrating out the radial information, ρ. That is,

$$\begin{aligned} R_f(\theta,\phi) &= \int_{\infty}^{\infty} T_f^2(\rho,\theta,\phi)d\rho \quad (2)\\ &= \int_{\rho}\Bigg[\sum_{j=1}^{J} f(x_j,y_j,t_j)\cdot \\ &\quad \delta(x_j cos\phi sin\theta + y_j sin\phi cos\theta + t_j cos\theta - \rho)\Bigg]^2 d\rho \end{aligned}$$

As mentioned earlier, it is not necessary to compute this transform on all pixels, since only a small (not necessarily unique) subset is required for recognizing activities in videos. As such, we only perform the R transform on values $\boldsymbol{x} \in SPIT$ (Spatio-Temporal-Interest-Points). That is,

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in SPIT \\ 0 & \text{else} \end{cases} \quad (3)$$

3.2. Bag of Visual Words

Now that we have a global descriptor of video content (*R* transform), we choose to characterize activity sequences using a more descriptive, discriminative, measure to account for intraclass variability. In this accord, we employ a BOVW representation using localized HOG3D features [4]. There are numerous reasons why a localized descriptor is more desirable than a global descriptor, as discussed in Section 3.3.

Local interest regions are first estimated using a spatio-temporal interest point detector by dense sampling of the video volume. In our implementation we use SURF detector to arrive at these STIP. However, we expect a 3D Harris detector to work better in this context as a SURF detector is essentially a scale invariant 2D interest point detector.

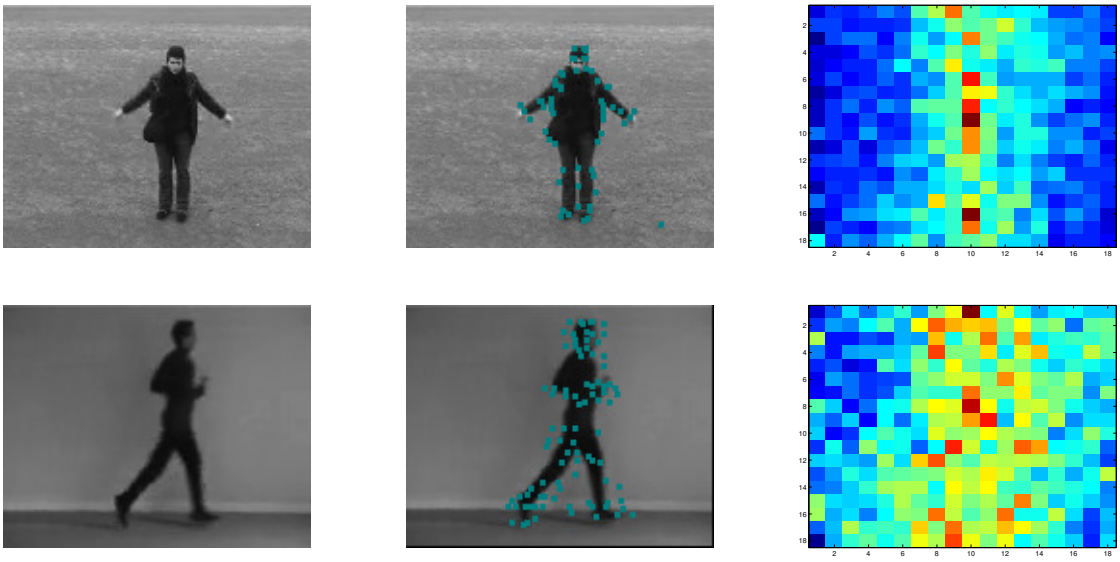


Figure 2. Sequence of frames showing the original frame, the detected spatio-temporal interest points and the 3D R Transform, for two different activities.

Once we have the STIP and their corresponding scales in time and space, we characterize each of these keypoints by extracting local HOG3D features. Specifically, this step involves generating a histogram of local 3D gradient vectors (including orientations) in a computationally efficient manner for different scales using spatio-temporal pyramids which are calculated using integral videos. This is similar to the concept of using integral images to smooth images, as ooposed to convolution with various multidimensional Gaussians.

HOG3D provides localized information in the 3D space around a keypoint (as specified by the time and space extents) by quantizing histogram information to planes existing on the faces of a dodecahedron or icosahedron. The resulting 3-dimensional histogram is vectorized to produce a consistent feature-vector descriptor. Several local HOG3D features from the training set collected and quantized using simple K-means clustering to form a codebook, termed the Bag of Visual Words (BOVW). Every keypoint in a video can then be matched with its closest Visual Words using a K-nearest neighbour search. Finally, each action sequence (video) is represented as a histogram with regard to all visual words, formulated as:

$$H = (n_1, n_2, \therefore, n_K) \tag{4}$$

where $n_i = \frac{N_i}{N}$ denotes the occurrence frequency of the i^{th} visual word in this video, and K is the number of visual words.

3.3. Feature Fusion

As noted previously, there may be considerable degeneracy in the spatio-temporal features extracted from a particular video. Intuitively, this can occur when salient-point patterns are similar (e.g. boxing vs sawing). As such, a robust recognition system requires a context-recognition subroutine, which first determines the type of video, or scenario, being analyzed prior to the particular activity. Certainly in this view, the context may be identified via an auxiliary feature appended to the original feature vector. In such a case, the classification tool would need to discriminate between a large number of classes efficiently. Even if a kernel approach is used, the distances between similar (but distinct) classes is not guaranteed.

Instead of relying on the precision of the classification tool, we apply a heuristic to the kernel to (1) identify the context and (2) perform efficient classification at one time. Our heuristic uses the *R* transform to identify the context of a video, and the BOVW to determine the specific class. The modified kernel is defined as follows:

$$K(i,j) = k_1(x_i^2, x_j^2) + k_2(x_i^2, x_j^2) + k_3(x_i^2, x_j^2) \quad (5)$$

$$k_1(x_i, x_j) = \sum_n min(x_i(n), x_j(n)) \quad (6)$$

$$k_2(x_i, x_j) = \frac{1}{k} \sum_{m \in N_j, m \neq j} k_1(x_j, x_m) \quad (7)$$

$$k_3(x_i, x_j) = \frac{1}{k} \sum_{m \in N_i, m \neq j} k_1(x_i, x_m) \quad (8)$$

$$N_m = knn(x_m^1, X_{train}, k) \quad (9)$$

where N_m refers to the k-nearest neighbors of feature vector m. In our work, we empirically choose $k = 10$. Furthermore, we should emphasize the notation used: x_i^s implies the s'th feature of index i is chosen. That is, we use the R-transform feature vector (x^1) for the context search via knn, and the BOWV representation (x^2) for the similarity measurement via the minimum overlap kernel $k_1(\cdot)$. Although an alternate kernel (such as the radial basis function) could be used, it is found empirically to be a poor kernel for embedding histogram information.

This heuristic kernel effectively weights the context of a video explicitly in the kernel function, rather than relying on a learned context via feature vectors. Moreover, the chosen context and class features are complementary. The BOVW based representations rely on the discriminative power of individual local cuboid features, while the R features exploit the global spatio-temporal distribution of the interest points. The benefit of such a feature-fusion method is apparent in its comparative performance (Section 4).

4. Results

In order to evaluate the performance of our descriptor, the experimental setup in this work closely follows the setup of Laptev et al. [5].

4.1. Dataset

The KTH actions dataset [7] contains six types of different human action classes: walking, jogging, running, boxing, waving, and clapping [7]. Each action class is performed several times by 25 subjects. The sequences were recorded in four different poses/scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. The background is homogeneous and static in most sequences. In total, the data consists of 2391 video samples. Following the experimental setup of [5], we initially divide the samples into training set (9 subjects) and test set (16 subjects).

4.2. Confusion Matrices

In order to evaluate the performance of our algorithm, we train six one versus all SVM's which combine to give a multi-class classifier. Tables 4.2 and 4.2 depict the results, in the form of Confusion Matrices, incorporating the R transform and BOVW features individually as discriminative features.

	1	2	3	4	5	6
1	21.88	35.94	40.62	0	1.56	0
2	22.22	26.98	50.80	0	0	0
3	12.50	12.50	75.00	0	0	0
4	15.62	15.62	12.50	1.56	29.69	25.00
5	14.06	12.50	10.94	4.69	34.38	23.44
6	12.50	12.50	15.62	0	12.50	46.88

Table 1. Confusion Matrix obtained by using only R features

	1	2	3	4	5	6
1	26.56	7.81	65.62	0	0	0
2	1.59	26.98	71.43	0	0	0
3	0	0	100	0	0	0
4	0	0	14.06	56.25	10.94	18.75
5	0	0	17.18	12.50	67.19	3.12
6	0	0	26.56	0	0	73.44

Table 2. Confusion Matrix obtained by using only BOVW features

	1	2	3	4	5	6
1	79.69	20.31	0	0	0	0
2	7.94	90.48	1.59	0	0	0
3	7.81	1.56	90.62	0	0	0
4	0	0	0	62.50	18.75	18.75
5	0	0	0	17.19	81.25	1.56
6	0	0	0	0	0	100.00

Table 3. Confusion Matrix obtained by using both R and BOVW features

In these matrices, [1,2,3,4,5,6] refers to the six activities in the KTH dataset namely - Boxing, Clapping, Waving, Jogging, Running, Walking.

Although individually the two features don't perform well, when combined using the feature fusion technique as described in Section 3.3, we get results which compare well with the results of the state-of-the-art methods. We calculate the average classification accuracy from the average diagonal elements of the confusion matrices; in this way we measure an average accuracy of **84.1%** on using the R feature for context calculation and the BOVW feature for context aware kernel calculation.

The following table positions our result in comparison with other well performing methods.

5. Conclusion

In this project, we implemented various salient point detectors and descriptors operating on a 3D spatio-temporal

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	89.0%	91.8%	80.9%	92.1%	–	–
Cuboids	90%	88.7%	82.3%	88.2%	89.1%	–
Hessian	84.6%	88.7%	77.7%	88.6%	–	81.4%
Dense	85.3%	86.1%	79.0%	88.0%	–	–
SURF	84.1%	–	–	–	–	–

Table 4. Comparison with the State of the art methods tested on KTH dataset.

volume and demonstrated the role of each component of our algorithm through experiments with multi-class SVM's. It is thus concluded that with better parameter selection and more descriptive features, it is possible to achieve a robust estimate of the action carried out by the subject in a video. Possible improvements in this model would aim at extending this method to activity recognition in the wild with multiple people in a single view. This method of activity recognition is also seen by the authors as directly complementing their previous efforts on video segmentation. A combined algorithm of both the aforementioned methods would be substantial milestone in generically characterizing videos.

References

[1] A. Averbuch and Y. Shkolnisky. 3d fourier based discrete radon transform. *Applied and Computational Harmonic Analysis*, 15(1):33–69, 2003.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[3] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

[4] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[5] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[6] T. Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

[7] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[8] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, pages 357–360. ACM, 2007.

[9] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008*, pages 650–663. Springer, 2008.

[10] C. Yuan, X. Li, W. Hu, H. Ling, and S. Maybank. 3d r transform on spatio-temporal interest points for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 724–730. IEEE, 2013.

[11] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238, 2007.