Recognizing Human Action from a Far Field of View

Chia-Chih Chen and J. K. Aggarwal
Computer & Vision Research Center / Department of ECE
The University of Texas at Austin
{ccchen|aggarwaljk}@mail.utexas.edu

Abstract

In this paper, we present a novel descriptor to characterize human action when it is being observed from a far field of view. Visual cues are usually sparse and vague under this scenario. An action sequence is divided into overlapped spatial-temporal volumes to make reliable and comprehensive use of the observed features. Within each volume, we represent successive poses by time series of Histogram of Oriented Gradients (HOG) and movements by time series of Histogram of Oriented Optical Flow (HOOF). Supervised Principle Component Analysis (SPCA) is applied to seek a subset of discriminantly informative principle components (PCs) to reduce the dimension of histogram vectors without loss of accuracy. The final action descriptor is formed by concatenating sequences of SPCA projected HOG and HOOF features.

A Support Vector Machine (SVM) classifier is trained to perform action classification. We evaluated our algorithm by testing it on one normal resolution and two low-resolution datasets, and compared our results with those of other reported methods. By using less than 1/5 the dimension a full-length descriptor, our method is able to achieve perfect accuracy on two of the datasets, and perform comparably to other methods on the third dataset.

1. Introduction

Recognition of human actions from a distant view is a challenging problem in computer vision. It is of significant interest in many applications, such as automated surveillance, aerial video analysis, sport video annotation and search. Various visual cues have been shown to be effective for representing human actions, including motion [8, 9], contours [3, 12], extremities [22], and body parts [5, 18], etc. Most of these features can be reliably extracted from image sequences of medium to high-resolution.

Similar to [8], our goal is to recognize actions from video sequences where human figures are less than 40 pixels in height. This is usually the case when actions are being imaged from a far field of view. Therefore, not only is the image resolution greatly reduced, but also the quality of visual cues is adversely effected due to turbulence. As shown in Figure 1(a), a person is waving both hands with optical flow vectors superimposed. The average width of his limbs is about 3 pixels, and the boundary between the body parts and background is vague. As a result, the computed optical flow is rather sparse and noisy. In our problem, we find that action classification with a single type of feature is easily subject to background noise and missing features. Moreover, there are certain human actions where one type of feature cannot fully capture their properties. For example, it is difficult to distinguish 'standing' from 'pointing' using optical flow alone. Therefore, instead of describing action by a single type of measure, we propose a novel descriptor which combines both human poses and motion information within a spatial-temporal volume.

We use Histogram of Oriented Gradients (HOG) to represent human poses. The HOG descriptor was originally proposed for human detection [7]. Due to its robustness, HOG has been successfully applied in the problem of action recognition [11, 16, 17, 21] and object recognition [4]. Similar to the gradient, optical flow is also a directional feature with magnitude. Therefore, we adopt the similar descriptor arrangement of HOG, and characterize human motion by Histogram of Oriented Optical Flow (HOOF).

To synthesize the action descriptor, sequences of HOG and HOOF features are extracted from overlapped space-time window of action frames. As in [8], we assume stabilized videos with human tracks are available to us. However, direct concatenation of the time series of both features will end up with a very lengthy descriptor vector. Hence we extend the technique of Supervised Principle Component Analysis (SPCA) [19] to perform feature selection based on the training data. Unlike regular PCA, SPCA aims at selecting a subset of PCs which best separate samples projected from different classes.

The major contribution of this work is two-fold. First we present a compact action descriptor which combines cues of human poses and motion. Our action descriptor is

shown to outperform similar descriptors which uses a single type of action feature, applies PCA for dimension reduction, or does not perform SPCA projection. Second, we extend SPCA to perform dimensionality reduction in a multiclass case. This step significantly speeds up the runtime of recognition without sacrificing accuracy. With the combination of radial basis function (RBF) kernel SVM, we achieve perfect accuracy on the Weizmann dataset [2] and our own low-resolution dataset called the Tower dataset. For another low-resolution dataset, the Soccer dataset [8], our performance is comparable to other tested methods.

This paper is organized as follows: Section 2 briefly reviews the related work. Our action descriptor and classification method are detailed in section 3. We discuss our experimental results in section 4, and conclude in section 5.

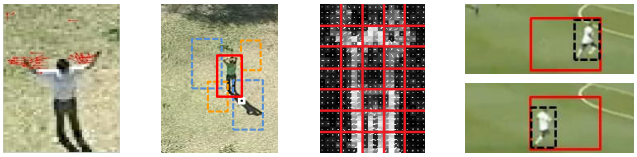
2. Related Work

The survey papers by Aggarwal and Cai [1], Gavrila [10], and Hu *et al*. [13] provide an extensive review of algorithms and systems for human tracking, motion analysis, action representation, and behavior recognition. In this section, we look at specifically the work which addresses the similar problem or adopts similar representation.

Efros *et al*. [8] propose an optical flow based motion descriptor for recognizing human action at a distance. Their descriptor is formed by rectified optical flow components in a spatio-temporal volume. They use *k*-nearest-neighbor classifier to perform action recognition and synthesis. As mentioned before, the use of motion feature alone is insufficient to characterize certain 'static' actions. Moreover, they compute the optical flow feature between figure-centric frames, which implicitly removes the velocity information of human movement.

In [17], Lu and Little employs the subspace projected HOG descriptor in a hybrid HMM classifier for the joint task of athlete tracking and action recognition. The space searched by PCA provides an efficient representation of the data, but it does not necessarily allow better separation of descriptor vectors from different actions.

Similar to our work, Ikizler *et al*. [14] use both human contour and motion features for action recognition. They characterize human contours by histograms of Hough transformed edges, and use coarse orientation bins to compute optical flow distribution. They train separate shape and motion classifiers and combine both classification results by averaging them. However, there is no evidence that shape and motion features are equally useful for distinguishing actions. Therefore, the linear combination of single feature trained classifiers may not be the optimal way of improving joint decision.



(a) (b) (c) (d)

Figure 1. (a) Motion feature presented in a far-field of view (b) given the track coordinate (white square) the bounding box for HOG extraction (red) is centered on the human figure by searching in the space of scale and translation (c) a human gradient map with our HOG geometry imposed (d) optical flow is computed between the union bounding boxes (red) of two consecutive frames.

3. Action Recognition

Our approach for recognizing action from a distant view video is outlined in Figure 2. In the following subsections, we first introduce the preprocessing step to acquire figure-centric frames. Then we briefly review the HOG and HOOF action features. From each feature space, we explain the method to select the top discriminative principle components. Finally, we present the action classifier.

3.1. Preprocessing and action features

Preprocessing. Given a stabilized video with tracks of human actors, the purpose of our preprocessing stage is to acquire figure-centric action sequences from the tracks. This step is critical, because in low-resolution video frames, even a minor misalignment of a bounding box can cause the loss of body parts or a large inclusion of background. To overcome this difficulty, we take the approach similar to [7] for human figure centralization. The major difference is that, instead of searching for all people in the entire frame, it is assumed that the person of interest is somewhere around the track coordinate. We train our figure centralization detector with HOG descriptors extracted from manually cropped figure-centric bounding boxes and negative samples from descriptors of patches around the figures. During runtime, within the neighborhood of interest, the detection window searches in the space of scale and translation (Figure 1.(b)). For a specific scale and translation which the SVM window classifier provides the highest probability estimate, the corresponding HOG vector and the window coordinates are stored. The recorded coordinates are then passed to the calculation of HOOF.

HOG. We use the HOG descriptor to characterize details of human poses. The essence of HOG is to describe local edge structure or appearance of object by local distribution of gradients [7]. Without directly using noisy gradient vectors as pixel-wise features, HOG gains robust representation

Figure 2. Flow diagram of our action recognition scheme. The focus of our method is in solid-line rectangles.

by presenting them as directional patterns over coarser spatial regions.

In HOG implementation, one action frame is divided into non-overlapping spatial grids (cells). For each pixel in the cell, we calculate its gradient vector $\mathbf{g}(x,y) = [g_x(x,y)\ \ g_y(x,y))]^{\mathrm{T}}$. The magnitude and orientation (four-quadrant tangent inverse) of a gradient vector are expressed as

$$m(x,y) = (g_x(x,y)^2 + g_y(x,y)^2)^{\frac{1}{2}} \quad (1)$$

$$\theta(x,y) = \tan^{-1}(g_y(x,y)/g_x(x,y)) \quad (2)$$

Based on $\theta(x,y)$ and (x,y), every $m(x,y)$ is weighted to vote for the nearest local orientation bins and also the adjacent cell histograms, respectively. Note that $\theta(x,y)$ should be insensitive to the order of contrast change, because the color variations in clothing and background do not provide extra information for the recognition task. To achieve this, $\theta(x,y)$ is further divided by the modulus π before binning. After accumulating the gradient histogram at each cell, for better invariance to the illumination changes, the concatenated histogram vector is normalized by the L2-norm. Figure 1(c) illustrates our HOG geometry.

HOOF. We characterize optical flow by the similar descriptor arrangement of HOG. In addition to the fact that both gradient and optical flow features are measured by 2D vectors, the accuracy of optical flow computation is very susceptible to the quality of image sequence. Therefore, in our scenario, representing optical flow by its local directional distribution is a more reliable option than using it by its exact value.

In the preprocessing step, we have already obtained the accurate estimates of bounding boxes which center on the human figures. Using this information, we are able to locate the minimum rectangular area which covers the moving person between two successive frames. As shown in Figure 1(d), the minimum rectangular area is in red and we name it union bounding box. Optical flow is then computed between pairs of successive union bounding boxes. Contrary to [8], we preserve the important information of human figure translations between frames. Therefore, computed optical flow vectors also carry the velocity information of human action. The procedure it takes to extract HOOF feature is the same as the major steps to compute HOG, except the use of the optical flow feature. We briefly review the important steps and explain the difference.

From the field of optical flow between two union bounding boxes, we extract vectors over the area covered by the first bounding box (dashed box, bottom frame of Figure 1(d)). The corresponding optical flow matrix is divided into non-overlapping spatial regions. We measure an optical flow vector by its magnitude $m_{of}(x,y)$ and orientation $\theta_{of}(x,y)$. In a spatial cell, every $m_{of}(x,y)$ is interpolated and aggregated into a local orientation histogram and the histograms nearby. The concatenated motion histogram is normalized to be more invariant to the scale of motion.

Similar to HOG feature, we need to take care of the issue with orientation mapping. In general applications, we do not use directions of actions as a cue to separate them. Therefore, a proper mapping of flow vectors is required so that different directions of the same action are treated as equivalent. The mapping is done by

$$\theta_{of} = \begin{cases} \mathrm{sgn}(\theta_{of}) \cdot (\pi - |\theta_{of}|), & |\theta_{of}| > \frac{\pi}{2} \\ \theta_{of}, & \text{otherwise} \end{cases} \quad (3)$$

By assuming that the profile view of human actions is being observed, this angular transformation makes motion representation symmetric about the vertical axis. However, there are applications where the direction of action is of interest. For example, in a soccer game, the player's action together with his/her motion of direction is usually considered as a whole. In this case, we can adjust the orientation mapping to meet the requirements accordingly.

3.2. Feature selection and action descriptor

Because of the high dimensionality of HOG and HOOF features in space-time, we perform dimensionality reduction for each type of the feature vectors before the final concatenation of the descriptor. In general, dimensionality reduction is carried out by feature extraction and selection. Classical approaches like PCA search for the directions which best represent the sample space. Even though the PCs found by PCA provide an efficient representation of the data, there is no evidence that the projected samples become more separable between classes.

The goal of SPCA is to select a subset of PCs which is most useful for discriminating data projected from different classes. In [19], the task is to detect sources of combustion from infrared imagery. In their binary class problem, PCs are first extracted from positive samples (sources of combustion). To evaluate the capability of a PC to distinguish

different classes of data, the discriminative value of a PC is defined as $d = \sigma^+/\sigma^-$, where σ^+ and σ^- are the standard deviation of the projected positive and negative samples, respectively. Therefore, the two classes of data are better separated in the space spanned by PCs with top d.

We extend SPCA to our multiclass action recognition problem. In the feature extraction step, for each action class i, the training samples are divided into $\mathbf{H}^i$ and $\mathbf{H}^{\forall-\{i\}}$ according to the labels. Here $\mathbf{H}^i$ denotes a n_f-by-n_i feature matrix where n_f is the length of feature vector and n_i is the number samples from class i. From the autocorrelation matrix of $\mathbf{H}^i$, we extract the matrix of principle components $\mathbf{PC}^i \in \mathbb{R}^{n_f \times n_f}$ by eigen value decomposition. The discriminative value of the j^{th} component (row) of $\mathbf{PC}^i$ is

$$d_j^i = \sigma_j^i / \sigma_j^{\forall-\{i\}} \tag{4}$$

$$\sigma_j^i = \sigma(\mathbf{PC}_j^i(\mathbf{H}^i - \bar{\mathbf{H}}^i)) \tag{5}$$

$$\sigma_j^{\forall-\{i\}} = \sigma(\mathbf{PC}_j^i(\mathbf{H}^{\forall-\{i\}} - \bar{\mathbf{H}}^i)) \tag{6}$$

and each column of $\bar{\mathbf{H}}^i$ is the mean vector of training samples from class i. In our implementation, we select the subset of PCs, $\mathbf{spc}^i$, of which the discriminative values of components are greater than one. Given a feature vector $\mathbf{h}$, its projection in the new space is

$$\widetilde{\mathbf{h}} = [(\mathbf{spc}^1(\mathbf{h} - \bar{\mathbf{h}}^1))^{\mathrm{T}} \ \ldots \ (\mathbf{spc}^{n_c}(\mathbf{h} - \bar{\mathbf{h}}^{n_c}))^{\mathrm{T}}]^{\mathrm{T}} \tag{7}$$

where $\bar{\mathbf{h}}^i$ is the mean vector of the samples in class i and n_c is the number of total action classes.

To characterize an action sequence, we divide the sequence into overlapped 'chunks' of frames, where each chunk is composed of sequential images of fixed duration. Time series of of HOG and HOOF features are extracted from every chunk of frames. After projecting them onto the corresponding subspaces, we denote each type of the transformed HOG and HOOF vectors by $\widetilde{\mathbf{h}}_g$ and $\widetilde{\mathbf{h}}_{of}$, respectively. The action descriptor extracted from frame $t+1$ to $t+N+1$ (covers N frames of optical flow field) is represented as

$$\mathbf{A} = [\widetilde{\mathbf{h}}_{g;t+1}^{\mathrm{T}} \ \ldots \ \widetilde{\mathbf{h}}_{g;t+N}^{\mathrm{T}} \ \widetilde{\mathbf{h}}_{of;t+1}^{\mathrm{T}} \ \ldots \ \widetilde{\mathbf{h}}_{of;t+N}^{\mathrm{T}}]^{\mathrm{T}} \tag{8}$$

which is further normalized by L2-norm before being employed by the classifier.

3.3. Action classification

To perform action classification, a multiclass SVM classifier is trained with labeled action descriptors. We adopt the implementation [6], of which the classifier prediction is made by a collection of one-against-one SVM classifiers. In the training phase, each binary SVM classifier leads to an inequality constrained quadratic optimization problem. We choose radial basis function (RBF) kernel for our SVM classifier because of the nonlinear relation between action classes and histogram features.

To estimate the best classifier for a dataset, grid search is performed in the space of parameter C and γ, where C is the weight of error penalty and γ determines the width of RBF kernel. The SVM classifier is decided by the set of (C, γ) which maximizes the cross-validation rate in the space of search. In the test phase, a preprocessed action sequence is segmented into intersected chunks of frames, where each chunk is characterized by an action descriptor. After SVM classification, descriptors are evaluated by the probability estimates of actions. We accumulate the probabilities over component descriptors, and classify the sequence as the action which gains the maximum votes.

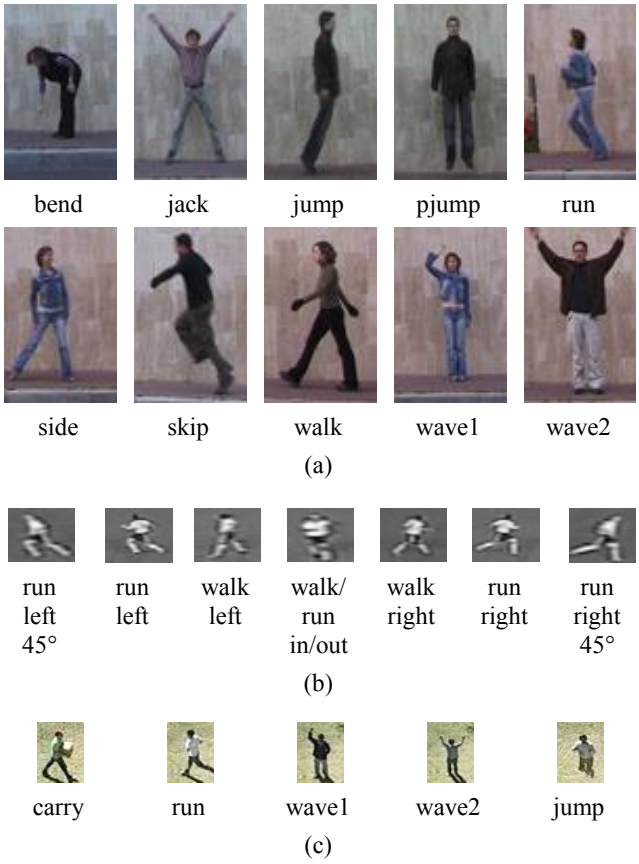


Figure 3. Sample frames from each action of (a) Weizmann dataset (b) Soccer dataset (c) Tower dataset.

4. Experimental Results

We have tested our method on three datasets, which include the normal resolution Weizmann dataset [2], the low-resolution Soccer dataset [8], and the low-resolution Tower dataset. We evaluate the performance on each dataset by

Method	Accuracy (%)
Our method	100
Fathi and Mori [9]	100
Blank *et al.* [2]	99.6
Jhuang *et al.* [15]	98.8
Hutan and Duygulu [11]	92.0

Table 1. Reported per-sequence accuracy on the Weizmann dataset.

leave-one-out cross validation, where one single action sequence is selected for testing at a time.

In general, good recognition results are achieved by setting the side of spatial cell to be the width of human limbs. The resolution of orientation bins is ranged from 10° to 20° depending on the dataset. To ensure the distribution of optical flow is not too sparse, we reduce the frame rate by half. Each chunk of frames covers 5 frames, and overlaps with the previous chunk by 4 frames.

Weizmann dataset. The Weizmann human action dataset contains 10 types of human actions performed by 9 different people. Every action is repeated 9 to 10 times so that there are 93 sequences in this dataset. The snapshots of action categories are shown in Figure 3(a). We use the provided foreground masks to extract human figures with fixed aspect ratio. Our method achieves 100% accuracy on this dataset. We list other reported results in Table 1 as a comparison.

Soccer dataset. The Soccer dataset is a low-resolution dataset collected by Efros *et al.* [8] from several minutes of World Cup soccer game. This dataset contains 66 action sequences from 8 classes. As shown in Figure 3(b), actions are distinguished by both action categories and the proceeding directions. Due to the high confusion between 'walk in/out' and 'run in/out', we treat them as the same action as in [9]. We also change the orientation mapping so that the in and out directions of the same action are recognized as the mirror of each other. Our performance and other reported per-descriptor accuracy on each action are presented in Table 2.

Besides the low-resolution video frames, the Soccer dataset poses other challenges to the recognition task. For example, in Figure 3(b), even a human observer may find it difficult to differentiate between 'run left' and 'run left 45°'. In addition, this dataset provides unstabilized figure-centric frames. Therefore, the computed optical flow does not contain the information of figure translation between frames. The unbalanced number of samples per class also reduces the classification accuracy on the minor classes. To alleviate these problems, we use background subtracted frames and randomly select the same number of descriptors from each class for training.

Except for 'run left/right 45°', our descriptor is comparable or better than other tested methods in Table 2. From the confusion matrix, substantial confusion occurs over the pairs of 'run left' versus 'run left 45°' and 'run right' versus 'run right 45°'. We assume that it is because of the nature of histogram representation, and speculate that histogram based descriptors may not be suitable for characterizing the subtle difference between the same type actions with large directional overlap. In most applications, it is expected that the action descriptor is general enough so that, for example, sequences of 'run left 45°' can be represented as the outliers of 'run left' or even 'run' class.

Action	Our Method	Efros [8]	Fathi [9]
run left 45°	0.47	0.67	0.63
run left	0.59	0.58	0.59
walk left	0.78	0.68	0.86
walk/run in/out	0.88	0.85	0.89
walk right	0.81	0.68	0.85
run right	0.58	0.58	0.65
run right 45°	0.52	0.66	0.53
Overall	0.66	0.67	0.71

Table 2. Comparison of descriptor level accuracy on each action of the Soccer dataset.

To verify our assumption, we combine the two pairs of actions which cause the most confusion and perform the experiment under the same settings. Table 3 shows the descriptor level confusion matrix when the number of classes is reduced to 5. Significant improvement is found over the the combined classes, while minor accuracy reduction is observed from the original actions due do the unbalanced number of samples per class after combination. Based on the class probabilities of the component descriptors of each sequence, the average accuracy per sequence is as high as 82.0%.

Tower dataset. To show the effectiveness of our method on more variety of human actions in low-resolution scenario, we created a dataset where human actions were being filmed from a distance. We name it the Tower dataset because it was taken from the top of a tower building. The Tower dataset contains 60 sequences of 5 different actions performed by 6 individuals. Figure 3(c) shows the sample frames from each action. In this dataset human figures are less than 40 pixels tall; therefore, trained with manually cropped figure-centric patches, the figure centralization detector is applied to ensure that each action frame is well centered on a figure. Following the similar settings of oriented histograms and space-time window, we obtain 100% accuracy on the Tower dataset as well.

To understand the representation effectiveness of different descriptor formats, we illustrate the corresponding ROC curves on downsampled versions of the Tower data. We have tested 4 combinations of action features and dimension

	run left/ run left 45°	walk left	run/walk in/out	walk right	run right/ run right 45°
run left/ run left 45°	***0.83***	0.08	0.01	0.01	0.07
walk left	0.12	**0.76**	0.09	0.02	0.01
run/walk in/out	0.01	0.07	**0.80**	0.07	0.04
walk right	0.01	0.03	0.09	**0.77**	0.11
run right/ run right 45°	0.04	0.01	0.03	0.12	***0.79***

Table 3. The descriptor level confusion matrix of the Soccer dataset when the number of classes is reduced to 5 (the overall accuracy is 78.66%).

reduction methods. They are denoted as PCA-HOG-[], []-HOG-HOOF, PCA-HOG-HOOF, and SPCA-HOG-HOOF. Here PCA-HOG-[] represents the PCA projected HOG descriptor, []-HOG-HOOF stands for the full-length joint feature descriptor, and PCA-HOG-HOOF is PCA projected HOG and HOOF time series. These descriptors all represent features in a spatio-temporal volume, and are employed by the same SVM classifier (parameters are optimized separately) to perform action recognition.

We perform 3-fold cross validation in a modified way to demonstrate descriptor performance on each action. That is we randomly select 4 sequences from total 12 sequences of each action for testing, and train on the labeled descriptors from the rest of the sequences. For each scale of the image resolution, we show only the ROC curves of action with the least area under the ROC curve (AUC). Figure 4(a) illustrates the comparison of all the 4 descriptors in the original resolution. Figure 4(b) and 4(c) correspond to the descriptor performance when frame resolutions are reduced to 36% and 16% of the original, respectively.

Our action recognition algorithm is implemented with MATLAB® and run on a Pentium 4 2.8GHz PC. Without further optimization, the average time required to classify a 10-descriptor sequence is ranged from 0.2 to 0.5 seconds. However, if we change the descriptor formation by neglecting the SPCA projection step, it takes 1.3 seconds on average. Because of the use of SVM classifier, the run time depends on the number of training samples [20].

5. Conclusions

When actions are being observed from a far field of view, available visual cues from human figures are usually sparse and vague. Therefore, action recognition algorithms that require an exact description of human shapes or motion may

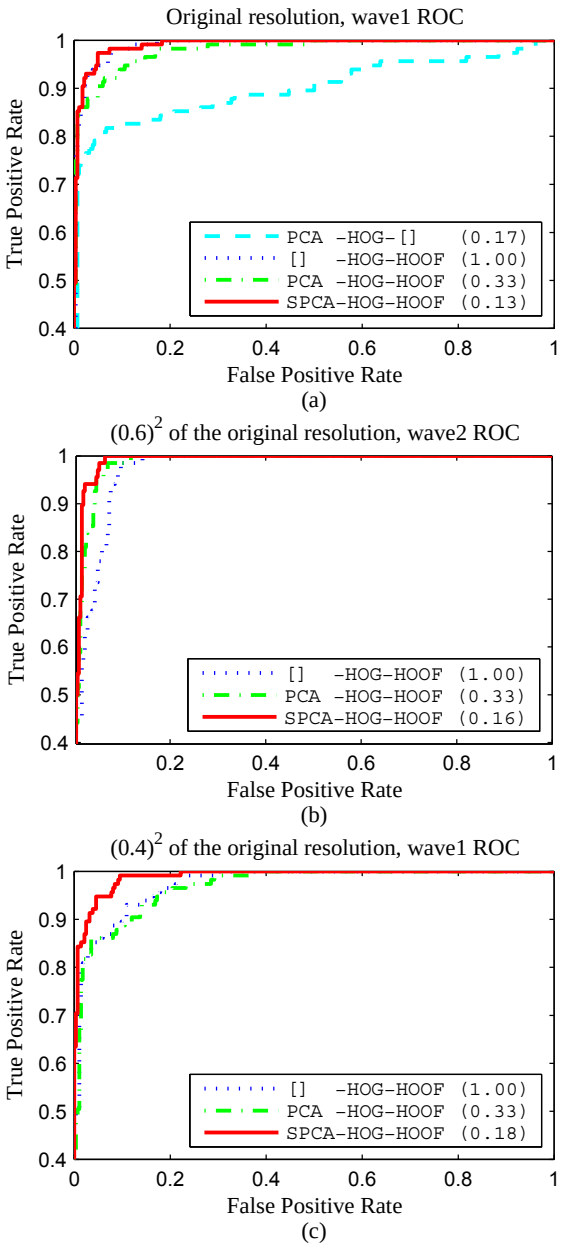


Figure 4. For the Tower dataset, we plot the one-against-rest ROC curve for the action with the minimum AUC. The performance of descriptors is evaluated when the frame resolution is (a) original, 40-pixel tall figures (b) 36% of the original, 25-pixel tall figures (c) 16% of the original, 15-pixel tall figures. The decimals in the parentheses represent the ratios of descriptor dimensions to the dimension of a full-length joint feature descriptor. In 4(a), the descriptor does not incorporate HOOF feature performs the worst. As shown in 4(b)(c), the ROC curves of the proposed SPCA-HOG-HOOF descriptor occupy the largest AUC in the lower resolution versions of the dataset. Note that as the frame resolution goes down, larger set of $\mathbf{spc}$ (Eq. (7)) is required from each class to provide better separation of projected samples.

suffer from the quantity as well as the quality of features. The proposed action descriptor is able to better accommodate these issues for two major reasons. First, the use of local orientation histograms to represent features is less susceptible to noisy data. Second, compared to a single feature descriptor, our descriptor is composed of two features so that it is more robust against low quality or loss of features.

Even though a human figure occupies much fewer pixels in a low-resolution video frame, the same amount of feature dimension is still required to characterize an action frame. In particular, our descriptor describes an action as a time series of poses and movements, which take considerable number of dimensions to represent. Moreover, blurry features in low-dimensional imagery add to the difficulty in distinguishing them. To reduce dimensionality while maintaining good accuracy, we extend an existing method to select a subspace of the transformed feature space that provides better separation of projected features for multiple classes.

In our experiments, our method achieves perfect accuracy on both the Weizmann dataset and the Tower dataset. We also show that the proposed action descriptor outperforms other formats of descriptor even when the resolution of figures is reduced to 16% of the original (Figure 4(c)). From the results on the Soccer dataset, it is shown that the velocity of the figure as a whole plays an important role in distinguishing directional actions.

6. Acknowledgement

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0135. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73:428–440, 1999. 2

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE International Conference on Computer Vision (ICCV)*, 2005. 2, 4, 5

[3] A. Bobick and J. Davis. The representation and recognition of action using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(3), 2001. 1

[4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. *Proceedings of the International Conference on Image and Video Retrieval*, 2007. 1

[5] B. Chakraborty, M. Pedersoli, and J. Gonzalez. View-invariant human action detection using component-wise hmm of body parts. *Lecture Notes In Computer Science*, 5098, 2008. 1

[6] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 4

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2005. 1, 2

[8] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *IEEE International Conference on Computer Vision (ICCV)*, 2003. 1, 2, 3, 4, 5

[9] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 5

[10] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73:82–98, 1999. 2

[11] K. Hatun and P. Duygulu. Pose sentences: A new representation for action recognition using sequence of pose words. *International Conference on Pattern Recognition (ICPR)*, 2008. 1, 5

[12] P.-C. Hsiao, C.-S. Chen, and L.-W. Chang. Human action recognition using temporal-state shape contexts. *International Conference on Pattern Recognition (ICPR)*, 2008. 1

[13] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34:334–352, 2004. 2

[14] N. Ikizler, R. G. Cinbis, and P. Duygulu. Human action recognition with line and flow histograms. *International Conference on Pattern Recognition (ICPR)*, 2008. 2

[15] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *IEEE International Conference on Computer Vision (ICCV)*, 2007. 5

[16] X. Li. Hmm based action recognition using oriented histograms of optical flow field. *Electronics Letters*, 2007. 1

[17] W. Lu and J. Little. Simultaneous tracking and action recognition using the pca-hog descriptor. *Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision (CRV)*, 2006. 1, 2

[18] M. S. Ryoo and J. K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. *International Journal of Computer Vision (IJCV)*, 2009. 1

[19] R. Santiago-Mozos, J. Leiva-Murillo, F. Perez-Cruz, and A. Artes-Rodriguez. Supervised-pca and svm classifiers for object detection in infrared images. *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2003. 1, 3

[20] I. Steinwart. Sparseness of support vector machines - some asymptotically sharp bounds. *In Proceedings of the 16th NIPS Conference*, pages 169–184, 2004. 6

[21] C. Thurau. Behavior histograms for action recognition and human detection. *Lecture Notes in Computer Science*, 4814:299–312, 2007. 1

[22] E. Yu and J. K. Aggarwal. Human action recognition with extremities as semantic posture representation. *International Workshop on Semantic Learning Applications in Multimedia in association with CVPR*, 2009. 1