High Performance Set of Features for Human Action Classification

S Brahnam[1] **and L. Nanni**[2]
[1]Computer Information Systems, Missouri State University, Springfield, MO, USA
[2]DEIS, IEIIT—CNR, Università di Bologna, Bologna, Italy

Abstract - *The most common method for handling human action classification is to determine a common set of optimal features and then apply a machine-learning algorithm to classify them. In this paper we explore combining sets of different features for training an ensemble using random subspace with a set of support vector machines. We propose two novel descriptors for this task domain: one based on Gabor filters and the other based on local binary patterns (LBPs). We then combine these two sets of features with the histogram of gradients. We obtain an accuracy of 97.8% using the 10-class Weizmann dataset and a 100% accuracy rate using the 9-class Weizmann dataset. These results are comparable with the state of the art. By combining sets of relatively simple descriptors it is possible to obtain results comparable to using more sophisticated approaches. Our simpler approach, however, offers the advantage of being less computationally expensive.*

Keywords: Human Action Classification; Local Binary Patterns; Gabor Filters; Histogram of Gradients; Machine Learning Techniques; Ensemble of Support Vector Machines.

1 Introduction

Human action classification can be defined as the task of matching videos containing human motion to a set of action class labels. This is a field of study that has only recently, within the last couple of years, been seriously investigated. Automatic labeling of action in video sequences has value in a variety of video searching applications, such as, locating various sports plays and dance moves in sports and music videos and suspicious behaviors (such as running out of a bank) in surveillance video [1]. There are also a number of artistic and gaming applications, as well as human-computer communication applications that could benefit from matching human motion to a set of action labels. For several general surveys of human action analysis that mention the importance of this problem see [2-6].

Automated human action classification is a difficult machine classification problem. Some challenges include large variations in action performance produced by variations in people's anatomy, problems with differences in recording setups and environmental changes (including lightening, camera viewpoint, and background complexity), and spatial and temporal variations (including variations in the rate people perform actions [7]).

Some significant research in human action analysis include Blank et al. [8], who used silhouettes to construct a space-time volume. In this study properties of the solution to the Poisson equation were utilized for activity recognition. In Kellokumpu et al., [9], a texture descriptor is used to characterize Motion History Images. In this study it is shown that a collection of local features can form a very robust description of human movement. In Boiman, and Irani [10] a new notion of similarity between signals is proposed. The regions of the "query" signal which can be composed using large contiguous chunks of data from the "reference" signal are considered to have high local similarity. Finally, in Ikizler, and Duygulu [11], a human pose, divided into rectangular patches based on their orientations, is represented by histogram of extracted.

Most of these earlier studies provide a number of datasets that were developed specifically to evaluate the systems reported in the various experiments: the KTH human motion dataset [12], which includes sequences of 25 actors performing 6 actions, the INRIA XMAS multi-view dataset [13], which contains 14 actions from 11 subjects captured from 5 viewpoints, the UCF sports dataset [14], the Hollywood human action dataset [15], a set of 8 actions performed by a variety of actors, and the Weizmann human action dataset [8], which contains recorded sequences of 10 actions from 10 actors. The Weizmann dataset, used in the studies reported in this paper, has become a widely used dataset for comparing action classification systems.

In this paper, we show that human action classification is best handled by combining multiple descriptors to boost performance. We combine three sets of features in order to obtain a reliable method for human action classification. In particular, we show that the response of Gabor filters and the standard application of the Local Binary Patterns to the mask images available in the Weizmann dataset obtains a >90% accuracy. The complete system, based on the combination of the features proposed in this paper with the histogram of gradients, obtains an accuracy of 97.8% using the 10-classes Weizmann dataset and a 100% of accuracy in the 9-classes Weizmann dataset. What our experiments highlight is the fact that fusion among simple feature extractors can obtain results comparable to the state of the art [9-11, 16]. We also show

that Local Binary Patterns (LBP) can be applied directly to the masks. Thus far only a variant of the LBP, Local Binary Patterns from Three Orthogonal Planes, has been tested for this problem [16]. Even though LBP-TOP has obtained better results with respect to our simple LBP, more frames must be considered in the feature extraction process using LBP-TOP, thereby increasing computational complexity. Because of this, LBP-TOP, unlike our method, could not function in a real-time system.

The remainder of this paper is outlined as follows. In section 2 we briefly describe our system architecture. In section 3 we provide a detailed description of the descriptors used in our experiments. In section 4 we provide experimental results. Finally, in section 5 we conclude this paper by noting contributions and offering a few suggestions for future research.

2 System architecture

In Figure 1 we provide a schematic of our complete system. We use the masked images available in the Weizmann[1] dataset [8, 17]. As mentioned above, this dataset is becoming a popular benchmark in this task domain. There are two different versions of the Weizmann dataset: the standard 10-class dataset and a reduced version that does not include the Skipping class (the 9-classes Weizmann dataset). The ten actions, illustrated in Figure 2, are performed by 9 subjects. In our system we resized the masks to 150 x 105. For more information on the dataset and masks, see [8, 17].

In the experiments reported in this paper, we combine the following: Gabor Filters, invariant local binary patterns, and histogram of oriented gradients. The descriptors are extracted from the mask images using a background subtraction algorithm (for the source code we used, see http:// maven. smith.edu/~nhowe/research/code/) [18]. The feature vector that describes a given sequence is obtained simply by summing the features extracted from each stand-alone frame. Since two of the descriptors we use are novel for this task, we provide a detailed description of feature extraction in section 3.

In the classification step a random subspace ensemble (RS) of support vector machines [19] is used. In the RS method each classifier is trained with a subset of all available features. We combine 50 linear support vector machines, each trained with 50% of all available features. For each set of features, a separate random subspace is trained. These three systems are then combined using the sum rule.

3 Human action classification descriptors

In this section we describe the descriptors used in our system: Gabor Filters, invariant local binary patterns, and histogram of oriented gradients.

3.1 Gabor filters

This descriptor is based on the well known FingerCode [20, 21] method developed for fingerprint matching. We named our approach for human action classification ActionCode.

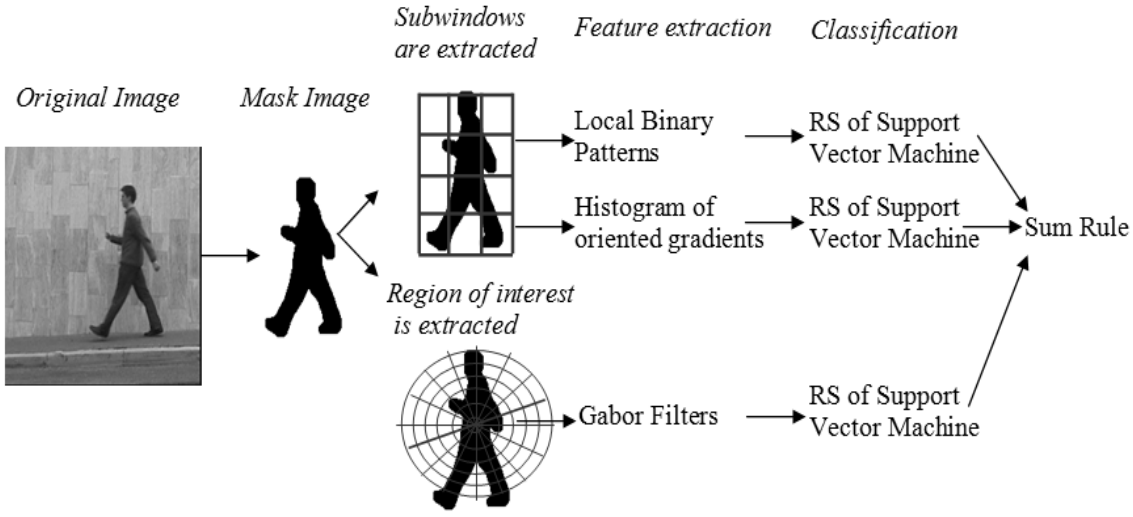

Figure 1. Proposed fusion system using new descriptors

[1] Available at http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

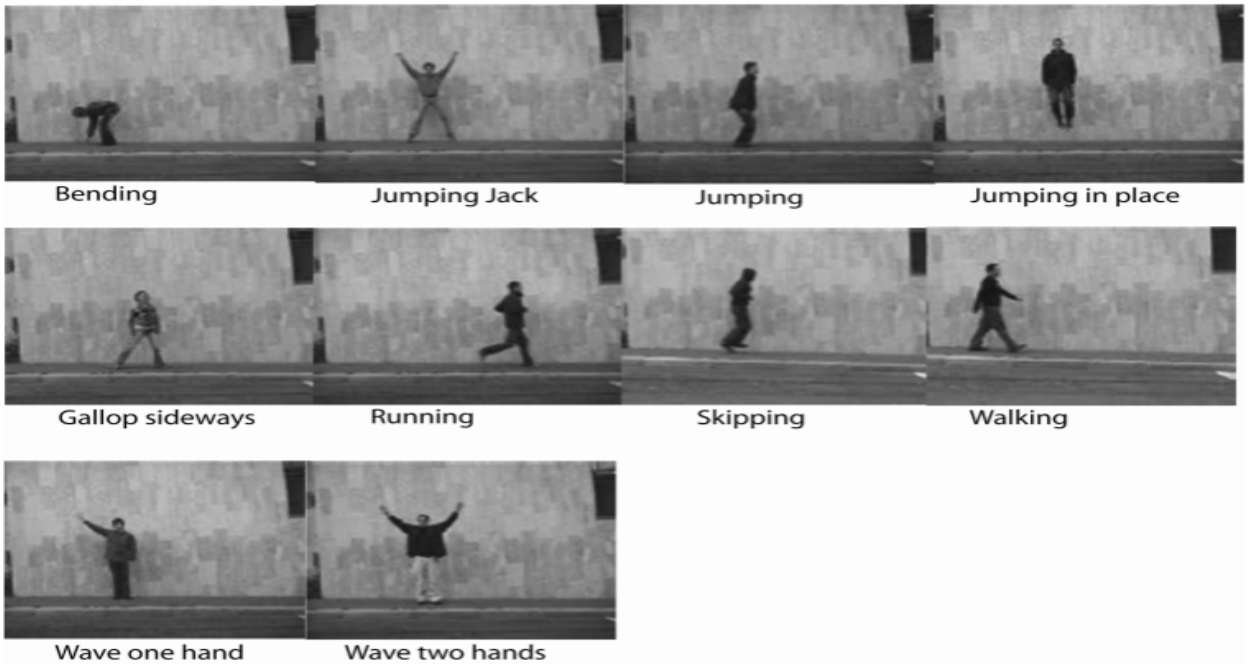

Figure 2. Some samples of the 10 action classes in the Weizmann dataset.

The basic ActionCode (AC) algorithm is as follows:

Step 1) Tessellate the region of interest around the center of the image[2];
Step 2) Filter the region of interest using a bank of Gabor filters;
Step 3) Compute the average absolute deviation from the mean of gray values in individual sectors in filtered images.

The region of interest (see Figure 3) is divided into 7 bands. In each band, 24 sectors are extracted (see [20] for details).

Figure 3. Region of interest in an image mask.

A symmetric Gabor filter has the following general form in the spatial domain [21]:

$$G(x, y; \nu, \sigma, \theta) = \exp\left(-\frac{x'^2 + y'^2}{2\sigma^2}\right) \cdot \cos(2\pi\nu x')$$

$$x' = x\sin\theta + y\cos\theta$$

$$y' = x\cos\theta - y\sin\theta$$

where ν is the frequency of the sinusoidal wave, θ is the orientation and σ is the standard deviation of the Gaussian envelope. In our experiments, the filters are obtained considering 12 angles (equally spaced between 0° and 180°).

3.2 Invariant Local Binary Patterns

This operator [22] has several properties of interest: it is low in computational complexity, and it is robust in terms of illumination changes and rotation invariant. The Local Binary Pattern is a histogram that is based on a statistical operator calculated by examining the joint distribution of gray scale values of a circularly symmetric neighborhood set of P pixels around a pixel **x** on a circle of radius R. In this study we use a multi-resolution descriptor that is obtained by concatenating three histograms calculated with the following parameters: (P=8; R=1) and (P=16; R=2). Each mask image is divided into 5×6 equal non-overlapping regions,[3] in each subregion the histograms are calculated, and then these 5×6=30 histograms are concatenated.

[2] It is supposed that the mask images are aligned.

[3] We have used these values because they are the default values offered in Poppe's Matlab code [7] and offer comparison.

3.3 Histogram of oriented gradients

The histogram of oriented gradients (HOG) was first proposed by Dalal and Triggs [23] as an image descriptor for localizing pedestrians. In this work we use weighted HOGs as implemented in [1], where the subregions are obtained by dividing each image cell into 5×6 equal non-overlapping regions.[3] In each subregion the orientation and magnitude of each pixel is calculated. The absolute orientations are discretized over 9 equally sized bins in the 0°- 180° range, and the resulting 9-bin histogram is calculated weighting each pixel by the magnitude of its orientation according to the histogram bin.

4 Experimental results

In Tables 1-2 we report our results and compare them with the state of the art. The method named *Fusion* is the combination by sum rule of the three systems tested. Our classification results using fusion (97.8% for the 10 class and 100% for the 9-class Weizmann dataset) matches the best performing systems reported to date using the same datasets [9-11, 16]. This demonstrates that simple feature extractors can obtain results comparable to the state of the art. Of course, these results are merely preliminary since only the Weizmann dataset is tried.

Table 1. Results on the 10-classes Weizmann dataset.

Method	**AC**	**LBP**	**HOG**
Stand-alone descriptor	90.3%	91.4%	94.6%
Fusion	97.8%		
LBP-Top [16]	95.6%		
Kellpkumpu et al [9]	97.8%		

These experimental results confirm what is well known in the machine learning community, namely, that combining methods is a simple approach for improving performance (see e.g., [24] in biometrics and [25] in bioinformatics).

Finally, in Table 3 we report the computational time for the extraction of each descriptor. These results are obtained using MATLAB 7.5 on a 2 GhZ Dual Core. Notice that the times are obtained considering a single frame. These results show that both LBP and HOG can be extracted in real time. In our opinion it is possible to obtain real time computation of AC using GPU.[4]

Table 2. Results on the 9 classes Weizmann dataset.

Method	**AC**	**LBP**	**HOG**
Stand-alone descriptor	95.2%	92.8%	96.6%
Fusion	100.0%		
LBP-Top [16]	98.7%		
Boiman and Irani [10]	97.8%		
Ikizler and Duyguku [11]	100.0%		

Table 3. Computational time

	AC	**LBP**	**HOG**
Computational Time	0.33 s	0.05 s	0.005 s

5 Conclusions

This paper focused on the study of descriptors for training an ensemble of machine learning algorithms for human action classification. We propose combining three relatively simple feature extractors for obtaining a system that performs as well as more complex systems. The ensemble proposed in this work has been tested on the Weizmann dataset, which is one of the most widely used benchmarks for comparing human action classification approaches. Our fusion results of 97.8% accuracy in the 10-class Weizmann dataset and a 100% accuracy in the 9-class Weizmann dataset performed as well as the best performance reported thus far.

This study makes a number of contributions. This is the first study to use fusion for human action classification. In addition, we introduce two new descriptors for this task: one based on the response of Gabor filters and the other based on the standard application of the Local Binary Patterns to the mask images. Although classification using these two descriptors (without combining descriptors) do not compare as well with the state of the art, they could be combined with other systems in order to further improve performance. Finally, our system of combining simple descriptors in fusion compares as well as more sophisticated systems but has the advantage of being computationally less intensive. It is very likely that our system could be used in a real-time system.

In future studies we plan on testing other fusion methods, in particular weighted approaches, where each method has different weights. In this way the best performing approaches can be given more weight in the classification combination step. We also want to test our methods using some of the other datasets mentioned in the introduction.

[4] Now a full GPU engine for MATLAB built on NVIDIA's CUDA technology named Jacket is available http://www.accelereyes.com/

6 Acknowledgements

This work was funded in part by Missouri State University Futures Grant: AI-ARTISTIC PROC 6-2011. The authors would like to thank Vonda Yarberry and Ruth Barnes, members of the AI-ARTISTIC PROC project, for their support. We would also like to thank Ronald Poppe at the Human Media Interaction Group, Department of Computer Science, University of Twente for sharing the Matlab code for the histogram of gradients. The Matlab code of the LBP is available at http://www.ee.oulu.fi/mvg/page/lbp_matlab.

7 References

[1] R. Poppe, "Evaluating example-based pose estimation: Experiments on the humaneva sets," in CVPR 2nd Workshop on Evaluation of Articulated Human Motion and Pose Estimation, 2007.

[2] Jake K. Aggarwal, and Qin Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding,* vol. 73, no. 3, pp. 428-440, 1999.

[3] Aaron F. Bobick, "Movement, activity and action: The role of knowledge in the perception of motion," *Philological Transactions of the Royal Society B: Biological Sciences,* vol. 352, no. 1358, pp. 1257-1265, 1997.

[4] Dariu M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding,* vol. 73, no. 1, pp. 82-92, 1999.

[5] Volker Krüger, Danica Kragic, Ašes Ide *et al.*, "The meaning of action: A review of action recognition and mapping," *Advanced Robotics,* vol. 21, no. 13, pp. 1473-1501, 2007.

[6] Antonius Oikonomopoulos, Maja Pantic, and Ioannis Patras, "B-spline polynomial descriptors for human activity recognition," in Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysi, Anchorage, AK, 2008, pp. 1-8.

[7] R. Poppe, "Discriminative vision-based recovery and recognition of human motion," Human Media Interaction, University of Twente, Tilburg, The Netherlands, 2009.

[8] M. Blank, L. Gorelick, E. Shechtman *et al.*, "Actions as space-time shapes," in International conference on computer vision, Beijing, China, 2005, pp. 1395-1402.

[9] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Texture based description of movements for activity analysis," in VISAPP, 2008, pp. 206 – 213.

[10] O. Boiman, and M. Irani, "Similarity by composition," in Neural Information Processing Systems (NIPS), 2006.

[11] N. Ikizler, and P. Duygulu, "Human action recognition using distribution of oriented rectangular patches," in ICCV workshop on Human Motion Understanding, Modeling, Capture and Animation, 2007.

[12] Christian Schüldt, Ivan Laptev, and Caputo Barbara, "Recognizing human actions: A local svm approach," in International conference on pattern recognition, Cambridge, United Kingdom, 2004, pp. 32-36.

[13] Daniel Weinland, Remi Ronfard, and Edmond Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding,* vol. 104, no. 2-3, pp. 249-257, 2006.

[14] Mikel D. Rodriquez, Javed Ahmed, and Mubarak Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8.

[15] Ivan Laptev, Marcin Marszalek, Cordelia Schmid *et al.*, "Learning realistic human action from movies," in Computer vision and Pattern Recognition, Anchorage, AK, 2008.

[16] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Human activity recognition using a dynamic texture based method," in BMVC08, 2008.

[17] L. Gorelick, Moshe Blank, Eli Shechtman *et al.*, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 12, pp. 2247-2253, 2007.

[18] N. Howe, and A. Deschamps, *Better foreground segmentation through graph cuts,* arXiv.org Tech Report arXiv:cs/0401017v2, 2004.

[19] R. O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed., New York: Wiley, 2000.

[20] A. K. Jain, S. Prabhakar, L. Hong *et al.*, "Filterbank-based fingerprint matching," *IEEE Transactions on Image Processing,* vol. 9, no. 5, pp. 846 – 859, 2000.

[21] L. Nanni, and A.Lumini, "Two-class fingerprint matcher," *Pattern Recognition,* vol. 39, no. 4, pp. 714-716, 2006.

[22] Timo Ojala, Matti Pietikainen, and Topi Maeenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Ieee transactions on pattern analysis and machine intelligence,* vol. 24, no. 7, pp. 971-987, 2002.

[23] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in 9th European Conference on Computer Vision, San Diego, CA, 2005.

[24] D. Maio, and L. Nanni, "Multihashing, human authentication featuring biometrics data and tokenised random number: A case study " *NeuroComputing,* vol. 69, no. December, pp. 242-249 2005.

[25] L. Nanni, and A. Lumini, "Ensemblator: An ensemble of classifiers for reliable classification of biological data," *Pattern Recognition Letters,* vol. 28, no. 5, pp. 622-630, 2007.