Simultaneous Human Detection and Action Recognition Employing 2DPCA-HOG

Mohamed A. Naiel(1) Moataz M. Abdelwahab(1) M. Elsaban(2) Wasfy B. Mikhael(3)
Fellow IEEE

(1)Nile University
6th October, Egypt

(2) Cairo Microsoft Innovation Lab
Cairo, Egypt

(3) University of Central Florida
Orlando, FL, USA

mohamed.naiel@nileu.edu.eg, mabdelwahab@nileuniversity.edu.eg, mikhael@mail.ucf.edu

***Abstract*— In this paper a novel algorithm for Human detection and action recognition in videos is presented. The algorithm is based on Two-Dimensional Principal Components Analysis (2DPCA) applied to Histogram of Oriented Gradients (HOG). Due to simultaneous Human detection and action recognition employing the same algorithm, the computational complexity is reduced to a great deal. Experimental results applied to public datasets confirm these excellent properties compared to most recent methods.**

I. INTRODUCTION

Human detection/action recognition, with its many applications such as video surveillance is one of the most challenging research problems in computer vision. Unfortunately, human classification faces many challenges including variation in appearance, clothes, poses, occlusion, shape, scale, illumination and environment.

Several approaches were used to solve this problem, a survey can be found in [1], and [2]. These approaches can be classified into body parts based approaches and single detection window approaches. For body parts based approaches, the human body is detected employing several detectors for the body parts. Mohan *et al.* [3] proposed component detectors for human body parts employing the Haar wavelet transform and used the SVM classifier to fuse each component score. In 2004 Mikolajczyk *et al.* [4] introduced several body parts detectors at multiple scales based on orientation features, and employing joint likelihood model which assembles these parts. In the classification stage, a coarse-to-fine cascade strategy was used which led to fast detection. Felzenszwalb *et al.* [5] performed a pyramid of HOG at multi-scales and employing a root filter of the whole body at near the top of the pyramid and part filters near the bottom of the pyramid, achieving excellent recognition accuracy. On the other hand, single window detection approaches extract low level features using one detector. For instance, In 2006 Dalal and Triggs [6] introduced a single window human detection algorithm with excellent detection results. This method uses a dense grid of Histograms of Oriented Gradients (HOG) for feature extraction and using linear Support Vector Machine (SVM) for classification. The HOG representation has several advantages. It captures edge or gradient structure that is the main feature of local shape, and it is robust for illumination changes, invariant for human clothes, and background changes. Recently Schwartz *et al.* [7] combined HOG features with texture and color information to generate more discriminative features, further Partial Least Squares (PLS) analysis was employed as a dimensionality reduction technique.

In 2004 the 2DPCA has been introduced to the facial recognition problem by Yang *et al.* [8]. In a previous contribution. Abdelwahab and Mikhael. [9] introduced the 2DPCA for human action recognition in videos using the 2D silhouettes extracted for humans in videos. This method reduces computational complexity by two order of magnitude, while maintaining high recognition accuracy and minimum storage requirements compared to existing methods.

In this paper, a novel algorithm for Human detection/action recognition is introduced, where 2DPCA is applied to Histogram of Oriented Gradients (HOG) represented into, *n* orientation bins, layers in 2D format. The main technical contributions of this work are two folds, first representing the HOG features in 2D format so that the relation between HOG features is maximized. Second, presenting a method that simultaneously perform human detection/ action recognition which reduces the computational time while maintaining comparable accuracy levels, as other state-of-the-art approaches.

This paper is organized as follows. Section 2 presents the overall system description,. Section 3 presents experimental results on a public datasets. Finally, conclusions are drawn in section 4.

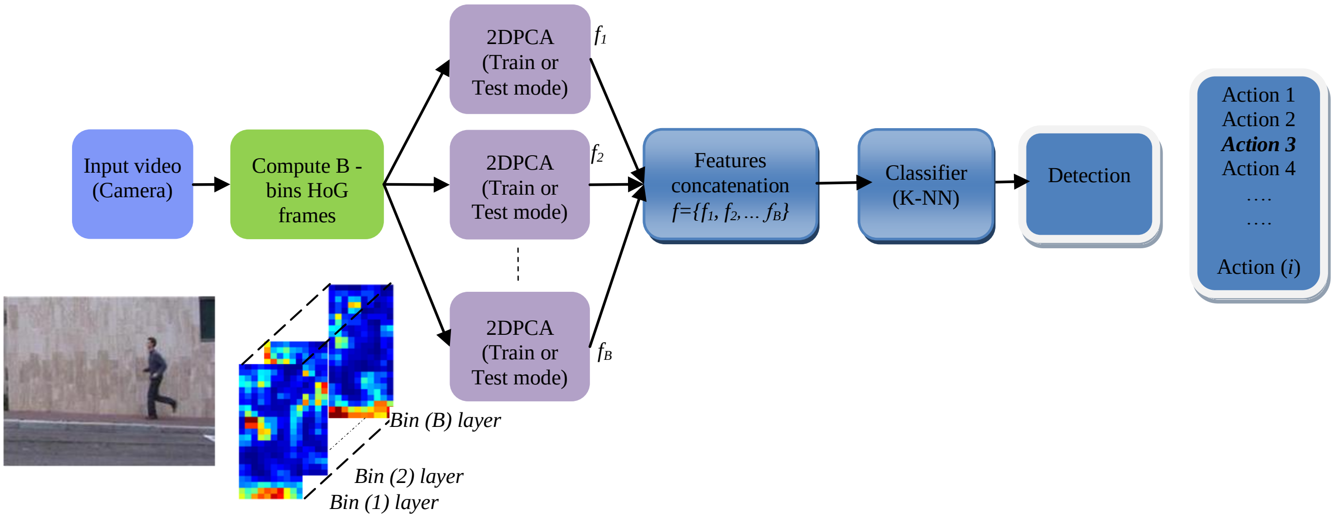
Input video (Camera)
Compute B - bins HoG frames
2DPCA (Train or Test mode)
2DPCA (Train or Test mode)
2DPCA (Train or Test mode)
f_1
f_2
f_B
Features concatenation $f=\{f_1, f_2, \ldots f_B\}$
Classifier (K-NN)
Detection
Action 1
Action 2
Action 3
Action 4
….
….
Action (i)
Bin (B) layer
Bin (2) layer
Bin (1) layer

Figure 1: The proposed human action recognition system.

II. Overall System Description

The proposed human detection/action recognition system consists of a parallel structure (*layers*) shown in Figure 1. First HoG *B*-bin features are extracted for every frame. The 2DPCA is applied on every bin layer independently, and then feature concatenation is used to fuse the 2DPCA features. The K-Nearest Neighbor (KNN) *Classifier* is used to infer the most likely class for the input view. Finally, this algorithm can be extended to multi-camera system using a majority voting technique to arbitrate between independent camera decisions. In the following subsections description for the HoG features representation and the proposed algorithm will be presented.

A. The HoG features

The proposed human descriptor is based on applying the 2DPCA on the HOG features, so that multiple instances can be extracted from the same frame. In our experiments the recognition system works as follows. Each detection window of size 128x64 pixels is divided into 15x7 blocks (with one cell overlap), where every block consists of 2x2 cells and each cell consists of 8x8 pixels. Each cell is represented by 9 Histogram of Orientation Gradients bins from 0° to 180°. These bins will be arranged into 9 images (or channels), where the spatial location is maintained. The output is 9 images each of size 30x14, compared to one feature vector of size 1x3780 in Dalal-Trigges [6] method (same numbers were used for comparison)

B. The Proposed Algorithm

Feature extraction is carried in the spatial domain using 2DPCA. The algorithm is divided into two modes, the training mode and the testing mode. The input patterns are the 2D-HoG *B*-bins frames. Let k denote the number of training frames, the j^{th} training frame is denoted by an (m x n) matrix $A_j(j = \{1, 2, \ldots, k\})$, and the average image of all training samples is denoted by $\bar{A}$. Let I denote the number of training videos, and Y denotes the number of frames for the i^{th} training video $N^i(i = \{1, 2, \ldots.., I\})$, where the size of every frame is (m x n). Let γ_{train} and γ_{test} denote the number of centroids per video computed in the training and testing modes, respectively. The training and testing algorithms are in Algorithms 1 and 2 respectively, where the training algorithm represents one path from the multi-camera system.

Training Algorithm
Input: $\{A_j\}_{j=1}^{k}, \{N^i\}_{i=1}^{I}, \gamma_{train}$
Output: matrices $V, C,$ and vector l

```
for $b := 1$ to $B$ step 1 do // for all HoG bins
    Compute the average for all frames in $\bar{A}^b = \sum_{j=1}^{k} A_j^b$;
    Compute the covariance matrix as
        $S^b = \frac{1}{k}\sum_{j=1}^{k}(A_j^b - \bar{A}^b)^T(A_j^b - \bar{A}^b)$;
    Compute the $r$ eigenvectors $\{\phi_q^b\}_{q=1}^{r}$ of $S^b$
        corresponding to the largest $r$ eigenvalues;
    $V^b \leftarrow [\phi_1^b, \phi_2^b, \ldots, \phi_r^b]$;
end for
for $i := 1$ to $I$ step 1 do // for all training videos
    for $y := 1$ to $Y$ step 1 do // for all frames in video
        for $b := 1$ to $B$ step 1 do // for all bins layers
            $F_y^{i,b} \leftarrow N_y^{i,b} V^b$;
            $f_y^{i,b} \leftarrow Concatenate\ (F_y^{i,b})$;
        end for
        $f_y^i \leftarrow \{f_y^{i,1}, f_y^{i,2}, \ldots, f_y^{i,B}\}$;
    end for
    Compute $\gamma_{\mathrm{train}}$ centroids for $i^{th}$ video as
        $C^i \leftarrow kmeans(f_y, \gamma_{train})$; and $l^i \leftarrow action\ label$;
end for
```

Algorithm 1: Training mode algorithm for every camera/multi-input

```
Testing Algorithm
Input: $\{N_j^t\}_{j=1}^{T}$, $V$, $C$, $l$, $Ncam$, $\gamma_{test}$
Output: $\hat{\theta}$
for $z := 1$ to $Ncam$ step 1 do// for every camera
    for $y := 1$ to $T$ step 1 do// for all testing frames
        for $b := 1$ to $B$ step 1 do // for all bins layers
            $F_y^{t,b} \leftarrow N_y^{t,b} V^b$;
            $f_y^{t,b} \leftarrow Concatenate\ (F_y^{t,b})$;
        end for
        $f_y^{t,i} \leftarrow \{f_y^{t,i,1}, f_y^{t,i,2}, \dots, f_y^{t,i,B}\}$;
    end for
    Compute $\gamma_{test}$ centroids for the testing video as $C^t \leftarrow kmeans(f_{\mathrm{y}}, \gamma_{test})$;
    for $y := 1$ to $\gamma_{test}$ step 1 do
        $\hat{\imath}_y \leftarrow argmin_{i\in\{1,2,\dots B\}} D_i\left(C_y^t, C^i\right)$; // using the Euclidean distance or any distance rule.
        $\hat{l}_y \leftarrow l^{i=\hat{\imath}_y}$; // decision per tested centroid
        $\widehat{D}_y \leftarrow D_{\mathrm{i}=\hat{\imath}_y}$; //minimum distance tested centroid
    end for
    if $\{\hat{l}_y\}_{y=1}^{\gamma_{test}}$ majority voting then
        $\hat{\theta}_z \leftarrow$ Action corresponding to majority labels.
        $\widehat{D}_z \leftarrow average\ \{\widehat{D}_p\}_{p\ \square\ \{y|\ \hat{l}_y=\hat{\theta}_z\}}$ ; //minimum distance per camera.
    else
        $\hat{\theta}_z \leftarrow$Action corresponding to minimum $\widehat{D}_y$.
        $\widehat{D}_z \leftarrow \min\ \{\widehat{D}_y\}_{y=1}^{\gamma_{test}}$ ; //minimum distance per camera.
    end if
end if
$\hat{\theta} \leftarrow$ Majority voting technique to infer the corresponding action for this view, where don't know decisions are ignored, if the majority voting is not satisfied, the system chooses the decision of the camera with minimum$\widehat{D}_z$.
```

Algorithm 2: Testing mode algorithm for the multi-camera system.

III. EXPERIMENTAL RESULTS AND ANALYSIS

The performance of the proposed algorithm was measured as follows, (i) For detection , the INIRIA person dataset was used [10]. (ii) For recognition, the Weizmann Action Dataset was employed [13]. All the results were obtained using CPU, 3 GHz dual core processor with 4GB of RAM .

3.1. INIRIA dataset

In this dataset persons are standing with different orientations, and with large variation in background views as shown in Figure 2. The training dataset consists of 1208 positive (each of size 96x160) and 1222 negative examples, where the positive examples are left-right reflected to give a total of 2416 positive examples. The testing dataset consists of 1126 (each of size 70x134) positive class and 453 negative class images. A random sample of six detection windows per image was taken for both training and testing negative datasets.

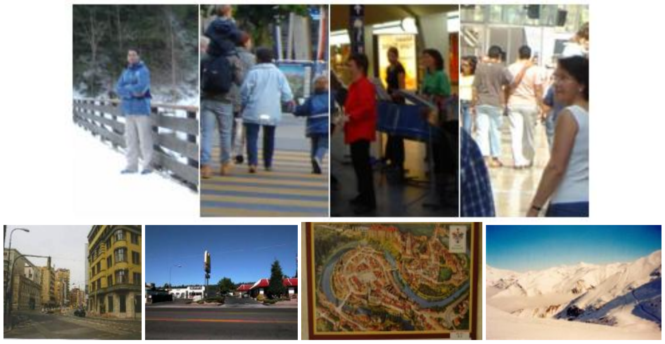

Figure 2. Sample images from INIRIA pedestrian dataset [10], shown examples for positive and negative classes, in the first and second row respectively.

In the proposed training algorithm the numbers of dominant eigenvectors (r) were selected for every channel to maintain 95% of the energy of the dominant eigenvalues. For every channel the system generates 2416 and 7332 feature vectors for the positive and negative classes respectively. The average time used for feature extraction from the 9-channels of the HOG with 2DPCA is 23.5 mins. For fast testing the K-means was employed to group the positive class into 30 centroids, while group the negative class into 250 centroids. The average time used for grouping was 15 mins. The average testing time per detection window (128x64) was 52 msec .

The 2DPCA recognition system was trained on the 2D Multilayer HOG images, and in the testing phase the multiple features were fused using voting technique. The AUC for our approach is 0.9841 while in [6] the AUC is 0.91. Also the Equal Error Rate (EER) for our system is 0.951 compared with Dollar *et al.* [12] 0.954, Maji *et al.* [11] 0.945.

Figure.3 compares the proposed algorithm performance with the best performance in previous published reports [11] and [12]. The average size of the feature vectors is 1x314 compared to 1x3780 for Dalal- Triggs [6]. Thus the feature space is reduced by 91.7%. Further grouping the features using K-means reduces the number of comparisons in the testing mode by a factor of approximately two orders of magnitude, while maintaining the highest recognition accuracy. These properties confirm that the proposed representation for the HOG features has discriminative properties. In other experiment, the parallel structure features generated from the 2DPCA were concatenated to give one feature vector for every input image in both the training and testing modes. This experiment gives consistent results with the parallel structure with voting scheme.

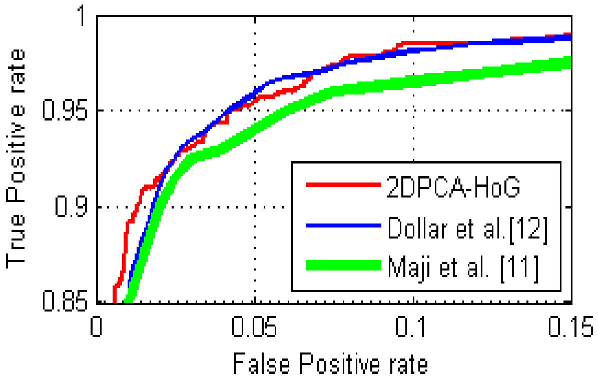

Figure 3. ROC Curve comparing the performance of the proposed algorithm with two recent reports [11] and [12].

3.2. Weizmann dataset

The proposed Algorithm was applied to the Weizmann Action Dataset [13] for human action recognition. Experimental results were compared to methods that were recently published. The Weizmann dataset consists of 90 low-resolution *(180 x 144, 50 fps)* video sequences showing nine different actors, each performing 10 natural actions such as walk, run, jump forward, gallop sideways, bend, wave one hand, wave two hands, jump in place, jump-jack, and skip, as

Table I shows that we are maintaining the excellent recognition accuracy compared with other recently published approaches. Table II presents our proposed algorithm run time, which is faster than other available record [10].

Method	**Accuracy**	**Testing technique**
HoG-9Bins/SD	*94.38%*	*Leave one-actor out*
Silhouette/SD [10]	*96.19%*	*Leave one-video out*
Silhouette/TD [10]	*96.75%*	*Leave one-actor out*
Ali & Shah [14]	95.75%	Leave one-actor out
Yuan *et al.* [15]	92.90%	Leave one-actor out
Yang *et al.* [16]	92.80%	Leave one-actor out
Niebles & Fei-Fei [17]	72.80%	Leave one-actor out

Table 1: Comparison of the average recognition accuracy on the Weizmann dataset with the best reported accuracy (Italic indicates our approach).

	Average time in sec	**Time standard deviation in sec**
Time HoG/Frame	0.4146	0.0120
Projection and Kmeans Time/Video	0.1295	0.0488
Classification time/Video	0.1190	0.0133
Total time/video	*25.1245*	*0.7821*

Table 2: Comparison of the average testing time on the Weizmann dataset

IV. CONCLUSIONS

A novel algorithm based on Two-Dimensional Principal Components Analysis applied to Histogram of Oriented Gradients for Human detection/Classification is presented. Experimental results applied to public datasets shows that the computational requirements have been reduced to a great deal while maintaining excellent recognition accuracy, compared to recent methods. As far as future work is concerned, using the proposed technique in the transform domain can lead to reduced computational and storage requirements.

V. REFERENCES

[1] D. M. Gavrila "The visual analysis of human movement: a survey" Journal of Computer Vision and Image Understanding, vol. 73, pp. 82-98, (1999).

[2] T. Moeslund, A. Hilton, and V. Kruger "A survey of advances in vision-based human motion capture and analysis", Journal Computer Vision and Image Understanding – vol. 104, no. 2, pp. 90-126, Nov.2006.

[3] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. IEEE Transactions on PAMI, vol. 23 no. 4, pp.349-361, 2001.

[4] K. Mikolajczyk, C. Schmid, and A. Zisserman."Human detection based on a probabilistic assembly of robust part detectors" in European Conference on Computer Vision 2004, Prague, Czech Republic, pages 69–82, May 2004.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan "Object detection with discriminatively trained part based models" IEEE Tran. on PAMI, vol.32, no.9, pp.1627-1645, Sep. 2010

[6] N. Dalal, and B. Triggs "Histograms of oriented gradients for human detection" IEEE Computer Vision and Pattern Recognition 2005, San Diego, CA, USA , pp.886-893 vol. 1, 25-25 Jun.2005.

[7] W. R. Schwartz, A. Kembhavi, D. Harwood, L. S. Davis, "Human detection using partial least squares analysis," IEEE International Conference on Computer Vision, Kyoto, Japan, pp.24-31, Sept.- Oct. 2009.

[8] J. Yang, D. Zhang, A. F. Frangi and J. Y. Yang "Two-dimensional PCA: A new approach to appearance-based face representation and recognition", IEEE Tran. on PAMI, vol.26, no.1, pp.131-137, Jan. 2004.

[9] M. A. Naiel, M. M. Abdelwahab, W. B. Mikhael "Human action recognition employing 2DPCA and VQ in the spatio-temporal domain", IEEE NEWCAS, Montreal, Canada, pp.381-384, Jun. 2010.

[10] http://pascal.inrialpes.fr/data/human/, last retrieved on Jan. 21, 2011.

[11] S. Maji, A.C. Berg "Max-margin additive classifiers for detection," IEEE International Conference on Computer Vision, Kyoto, Japan, pp.40-47, Sep.-Oct. 2009.

[12] P. Dollar, Zhuowen Tu; Hai Tao, S. Belongie, "Feature mining for image classification" IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, pp.1-8, Jun. 2007.

[13] www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html,

[14] Saad Ali, Mubarak Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," IEEE PAMI, vol. 32, no. 2, pp. 288-303, Feb. 2010.

[15] C. Yuan, Xi Li, W. Hu, H. Wang 'Human action recognition using pyramid vocabulary tree', ACCV, pp. 527-537,Sep. 2009.

[16] W. Yang, Y. Wang, and G. Mori, "Human action recognition from a single clip per action", 2nd MLVMA (at ICCV), Japan, September 2009.

[17] J. C. Niebles and L. Fei-Fei. "A hierarchical model of shape and appearance for human action classification", IEEE CVPR, Minneapolis, MN, pp. 1-8, June 2007.