Modeling Motion of Body Parts for Action Recognition

Khai N. Tran
khaitran@cs.uh.edu
Ioannis A. Kakadiaris
ioannisk@uh.edu
Shishir K. Shah
sshah@central.uh.edu

Department of Computer Science
University of Houston
4800 Calhoun Road
Houston, TX 77204 USA

Abstract

This paper presents a simple and computationally efficient framework for human action recognition based on modeling the motion of human body parts. Intuitively, a collective understanding of human part movements can lead to better understanding and representation of any human action. In this paper, we propose a generative representation of the motion of the human body parts to learn and classify human actions. The proposed representation combines the advantages of both local and global representations, encoding the relevant motion information as well as being robust to local appearance changes. Our work is motivated by the pictorial structures model and the framework of sparse representations for recognition. Human part movements are represented efficiently through quantization in the polar space. The key discrimination within each action is efficiently encoded by sparse representation to perform classification. The proposed method is evaluated on both the KTH and the UCF action datasets and the results are compared against other state-of-the-art methods.

1 Introduction

Human action recognition is a challenging problem that has received considerable attention from the computer vision community in recent years. Its applications are diverse, spanning from its use in activity understanding for intelligent surveillance systems to improving human-computer interactions. The challenges in solving this problem are multifold due to the complexity of human motions, the spatial and temporal variations exhibited due to differences in duration of different actions performed, and the changing spatial characteristics of the human form in performing each action [16, 30].

A number of approaches have been proposed to address these challenges over the past few years [12, 23]. Temporal templates have been proposed and used to categorize actions [5]. Methods that explicitly model relative changes in spatial descriptors over time [1], or estimates of global and local motion [10, 14] have also been used. Methods that use the silhouette of the body to construct more sophisticated representation for human action have been proposed [36]. More recently, spatio-temporal feature-based approaches have been proposed and demonstrated for various action recognition applications. Representations based

© 2011. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

BMVC 2011 http://dx.doi.org/10.5244/C.25.64

on a set of interest points are used to capture key variations in both space and time [8, 20, 24]. Space-time volumes built based on a global shape estimated by the human silhouette was proposed by Blank *et al.* [4]. Correlation of local features [29] and autocorrelation of features in space-time [18] have also been used to describe human movements.

Ideally, the desired representation for actions should generalize over variations in viewpoint, human appearance, and spatio-temporal changes. Moreover, the action representation must be sufficiently rich in its descriptors to allow for the robust recognition of actions. Human action representation can be divided into two categories: global representations and local representations [25]. The global representations can encode much of the information but they are more sensitive to the environment (e.g., viewpoint, noise). Local representations are less sensitive to the environment but they depend on the accuracy of interest point detectors. Robust action representation should encode most of the descriptive and discriminative information while being less sensitive to the environment.

Recently, pictorial structure models have been used for encoding spatial variations in geometric structures. The use of these models has been demonstrated for robust detection of object parts as well as for part-based recognition of articulated objects [3, 11, 27]. In the context of human action recognition, it is reasonable to assume that each action can be described as a combination of movements of different body parts (e.g, head, hands, legs and feet). Intuitively, a collective understanding of human part movements can lead to better understanding and representation of any human action. In this paper, we propose a generative representation of the motion of human body parts to learn and classify human actions. The proposed representation combines the advantages of both local and global representations, encoding the relevant motion information as well as being robust to local appearance changes. Our representation is motivated by the recent success of pictorial structures [3, 11, 27] and our recognition framework is inspired by the fundamentals of sparse representation for recognition [32]. The contributions of our work are:

1. *A representation of human action using a combination of body part movements.* We propose a new representation described by a combination of human body-part movements corresponding to a particular action. We use the polar space to represent the pattern of each human body-part movement.

2. *Sparse representation-based recognition for human actions using the proposed human body-part representation.* We propose a computationally efficient algorithm capable of discriminating the key differences in movement of each body-part pertaining to a particular action, thereby providing a robust and accurate recognition of complex actions.

The rest of the paper is organized as follows. Section 2 describes the human action representation proposed and used in this work along with the classification algorithm used to address the recognition task. Experimental results and evaluations are presented in Section 3. Finally, Section 4 concludes the paper.

2 Approach

2.1 Human Action Representation

A person is represented as a collection of body parts that are connected together in a deformable configuration. A common way to express such a part-based model is in terms of an

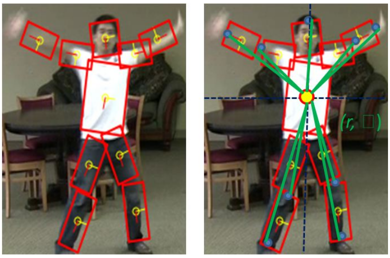

Figure 1: Depiction of human body part extraction and transformation to a local coordinate system.

undirected graph $G=(V,E)$, where the vertices $V=\{v_1,...,v_P\}$ correspond to the P body parts, and there is an edge $(v_i,v_j)\in E$ for each pair of connected part v_i and v_j. A part-based person representation is given by a configuration of P parts $L=\{l_1,...,l_P\}$ where, l_i specifies the location of part v_i. The location of each part is its location in the image space. Pertaining to human action recognition, each video frame provides us a snapshot of the action in terms of a configuration of the P body parts. Thus, a human action can be represented as a sequence of configurations of its body parts.

The configuration encodes the geometry of the human body or relationships between the body parts in the image space. As the action progresses through the video, the changes in the locations of the human body parts leads to changes in the configurations in the corresponding frames. To incorporate the relative changes in the body part locations, which collectively constitute the human action, we need to transform the human part locations from the image coordinate space to a local coordinate system. We consider the center of the torso as the origin of the local coordinate system. Thus, location l_i of the part v_i in the image coordinate space will become $l_i^T=\{x_i,y_i\}$ in the new local coordinate system. Using polar mapping, the location of the part will be represented by two components (r_i,θ_i) in the new coordinate space, where $r_i=\sqrt{x_i^2+y_i^2}$ is the distance of part v_i from the center of the torso and $\theta_i=\arctan(\frac{x_i}{y_i})$ is the orientation of the part v_i with respect to the vertical axis. As a result, a snapshot of a human action can be represented by the set of P body part location vectors (r_i,θ_i). Figure 1 shows a pictorial representation of the part-based model and the transformation to a local coordinate space. It is possible that human actions can be distinguished based on distinct motion patterns of even a single body part. For example, when waving with two hands and waving with one hand, the motion patterns of almost all body parts are the same, except motion of one hand. This constitutes a key difference that can be used to differentiate the two actions. As an example, Figure 2 shows the distribution of θ and r for two body parts across four actions (Diving, Golf-Swing, Kicking, and Lifting). The distributions are distinct enough to be used for recognizing and can thus, be used as a robust, compact, and highly discriminative part motion descriptor. Our key contribution is the novel use of 2D histograms of body part locations in polar geometry as a representation of human action. Collectively, all the 2D histograms of each body part's location generated over the entire video form a description of relative motion of body parts that constitutes a specific action. Such a representation is rich in human action description, since it encodes the meaningful changes in the relative positions of body parts over time. Each body part's motion descriptor is a 2D histogram of size $R\times O$, where R and O are the numbers of radial and orientation bins, respectively. The 2D histogram is treated as a $R\times O$ motion descriptor

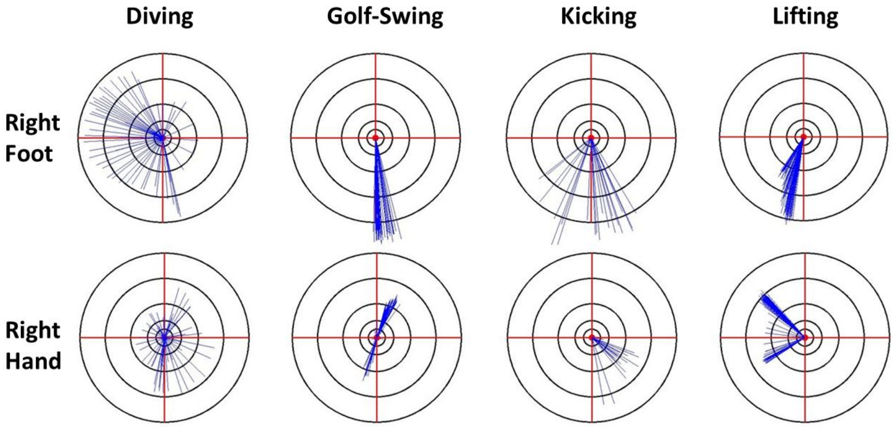

Figure 2: Depiction of the motion patterns of two human body parts for different human actions.

image. Thus, every human action is represented by P motion descriptor images, each describing the motion of the P body parts. Figure 3 depicts a representation of the body part motion descriptors, where the pixel value at each location in the motion descriptor image is represented by the number of times the respective body part is observed at that orientation.

2.2 Sparse Representation-based Recognition

Using the above action representation, an action recognition task can be considered as a 2D recognition problem. The dimensionality of the recognition problem is naturally managed since the motion representation efficiently encodes the temporal evolution of the action. This significantly reduces the recognition complexity.

In this section, we introduce our action recognition algorithm inspired by the sparse representation-based recognition algorithm [32]. A test sample comprised of P motion descriptors can be represented by a complete dictionary whose base elements are the training samples themselves. In other words, for each body part motion descriptor, the test sample can be represented as a linear combination of just those training samples that belongs to the same class. In mathematical terms, this representation is naturally sparse and the sparsest linear representation of the test sample can be recovered efficiently via ℓ_1-minimization.

Let us define the set of K human action classes to be recognized. A basic problem in action recognition is to determine the class that a new test sample belongs to. Let us consider a set of n_k training videos for the k^{th} action class where each video results in P motion descriptor images. As a result, for a particular human body part p, we will have a set of n_k training samples from the k^{th} class as columns of a matrix $A_k^p = [a_{k,1}^p, ..., a_{k,n_k}^p] \in \mathbb{R}^{m \times n_k}$, where each motion descriptor image of part p is identified as the vector $a^p \in \mathbb{R}^m (m = R \times O)$.

Let us consider the testing video y resulting in P motion descriptor images which are identified as the set of P vectors $\{y^p \in \mathbb{R}^m \mid p = 1, ..., P\}$. For a particular human body part p, we want to represent a test sample y^p as a sparse linear combination of training samples. Given sufficient training samples of the k^{th} action class, $A_k^p = [a_{k,1}^p, ..., a_{k,n_k}^p] \in \mathbb{R}^{m \times n_k}$, any new test sample $y^p \in \mathbb{R}^m$ from the same class approximately lies in the linear span of the training samples associated with action class k; i.e.,

$$y^p = w_{k,1}^p a_{k,1}^p + ... + w_{k,n_k}^p a_{k,n_k}^p, \qquad p = 1, ..., P \tag{1}$$

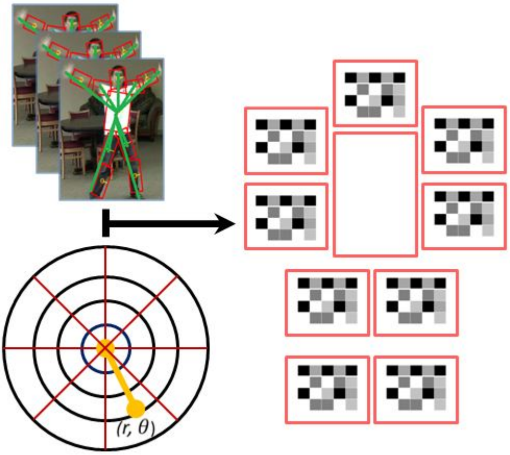

Figure 3: Human action representation by transforming the human body part motions to a polar histogram.

where $\{w^p_{k,j} \in \mathbb{R} \mid j = 1,...,n_k\}$.

Let us define a matrix A^p for the entire training set as the concatenation of the n training samples of all K action classes for particular human part p; i.e.,

$$A^p = [A^p_1,...,A^p_K] = [a^p_{1,1}, a^p_{1,2},..., a^p_{K,n_k}]. \qquad (2)$$

The linear representation of y^p can be rewritten over complete training samples as:

$$y^p = A^p x^p_0 \in \mathbb{R}^m, \qquad p = 1,...,P \qquad (3)$$

where $x^p_0 = [0,..0, w^p_{k,1}, w^p_{k,2},..., w^p_{k,n_k}, 0,...,0]^T \in \mathbb{R}^n (n = \sum_{k=1}^K n_k)$ is the coefficient vector whose elements are zero except for some that are associated with the k^{th} class.

This representation is naturally sparse if the number of action classes K is reasonably large. The more sparse is x^p_0, the easier it is to accurately determine the action class of test sample y^p for particular body part p. This motivates us to seek the sparsest solution to $y^p = A^p x^p$ by solving the following optimization problem:

$$(\ell_0): \qquad \hat{x}^p_0 = \arg\min \|x^p\|_0 \textbf{ subject to } y^p = A^p x^p \qquad (4)$$

where $\|x\|_0 = \#\{i : x_i \neq 0\}$.

However, the problem of seeking the sparsest solution of an underdetermined system of linear equations is NP-hard, and is difficult even to approximate [2]. Recent developments in the theory of sparse representation indicate that if the solution sought is sparse enough, the solution of the ℓ_0-minimization is equal to the solution to the following ℓ_1-minimization problem [9]

$$(\ell_1): \qquad \hat{x}^p_1 = \arg\min \|x^p\|_1 \textbf{ subject to } y^p = A^p x^p. \qquad (5)$$

The problem (ℓ_1) can be cast as a linear programming problem and solved using modern interior-point methods, simplex methods or other efficient techniques when the solution is known to be very sparse. Those methods can obtain the global solution of a well-defined optimization problem.

Given a new test motion descriptor y^p for particular part p, we first compute its representation $\hat{x}_1^p$ via Eq. 5. If this test sample belongs to action class k, ideally, the nonzero entries in the estimated $\hat{x}_1^p$ will be associated with multiple columns of A^p from a single action class k. However, due to noise and modeling error, the estimated $\hat{x}_1^p$ can contain small nonzero entries associated with multiple action classes. To handle this problem, we can classify y^p based on how well the coefficients associated with all the training samples of each class reproduce y^p. For each class k, let us define function $\phi_k^p : \mathbb{R}^n \rightarrow \mathbb{R}^n$ to be the characteristic function that selects the coefficients associated with the k^{th} class. For $x \in \mathbb{R}^n$, let $\phi_k^p(x) \in \mathbb{R}^n$ be the vector whose only nonzero entries associated with class k are kept from vector x. As a result, we can approximate the given test sample y^p as $\hat{y}_k^p = A^p \phi_k^p(\hat{x}_1^p)$. This allows us to obtain the residual between y^p and $\hat{y}_k^p$ for a particular human body part p, computed as:

$$r_k^p(y^p) = \|y^p - A^p \phi_k^p(\hat{x}_1^p)\|_2. \tag{6}$$

For a new test action y, which is characterized by P motion descriptors $\{y^p \mid p = 1, ..., P\}$ corresponding to P human body parts, we compute the total residuals of all P body parts over all K action classes. Then, we classify y to belong to the action class k that minimizes the total residual:

$$y \rightarrow k^* \text{ where } k^* = \arg\min_k \sum_{p=1}^{P} r_k^p(y^p). \tag{7}$$

The complete recognition procedure is summarized in Algorithm 1 below. Our implementation for ℓ_1-minimization problem is based on the model proposed by Wright *et al.* [33].

Algorithm 1 : Sparse Representation-based Recognition

```
Input: Data matrix $A = \{A^p \mid p = 1, ..., P\}$ where $A^p = [A_1^p, ..., A_K^p] \in \mathbb{R}^{m \times n}$ for $K$ action
classes. A test sample $y = \{y^p \in \mathbb{R}^m \mid p = 1, ..., P\}$.
for $p = 1$ to $P$ do
    Normalize the columns of $A^p$ to have unit $\ell_2$-norm
    Solve the $\ell_1$-minimization problem:
        $\hat{x}_1^p = \arg\min \|x^p\|_1$ subject to $y^p = A^p x^p$        (8)
    for $k = 1$ to $K$ do
        Compute the residuals $r_k^p(y^p) = \|y^p - A^p \phi_k^p(\hat{x}_1^p)\|_2$
    end for
end for
Output: Assign $y$ to class $k^*$ where $k^* = \arg\min_k \sum_{p=1}^{P} r_k^p(y^p)$.
```

3 Experiments and Results

In this section, we describe selected experiments that demonstrate the performance of the proposed algorithm for human action recognition. We test our algorithm using two benchmark datasets: the KTH human motion dataset [28], and the UCF Sport dataset comprising of 200 video sequences [26]. The KTH dataset consists of six actions (Boxing, Hand-clapping, Hand-waving, Jogging, Running and Walking) performed several times by 25 subjects in different scenarios of outdoor and indoor environments with scale changes. The UCF Sport

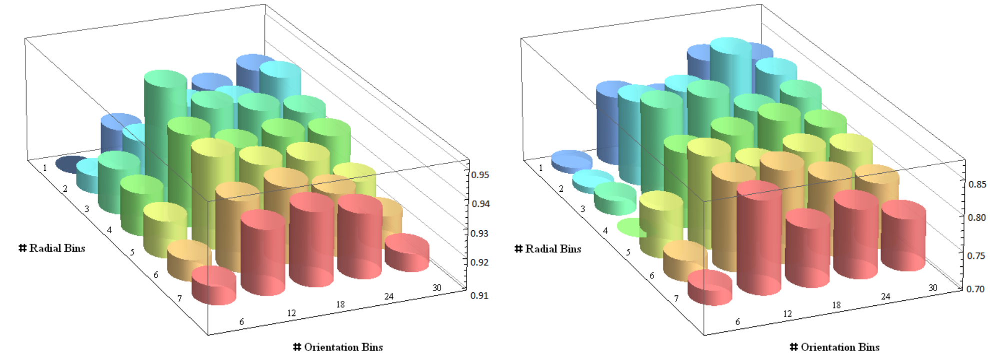

Figure 4: Recognition rates for different quantization levels of the radius and orientation: (L) on the KTH dataset and (R) on the UCF Sport dataset.

dataset consists of nine sports actions (Diving, Golf-Swinging, Kicking, Lifting, Horse-Riding, Running, Skating, Swinging, and Walking), which are captured in unconstrained environments with a wide range of scenes and viewpoints.

3.1 Dataset Preprocessing

In our implementation, we consider $P = 9$ body parts that contribute in expressing any human action. These body parts are: head, left and right upper and lower arms, left and right upper and lower legs. We use the pictorial structures model to identify human body parts in each frame of the video. First, we implement the entire framework in an automatic pipeline by using pictorial structure model, proposed in [3], to extract human body parts. Second, we manually annotate the video data to extract body parts in each of the video frames. This approach ensures consistent detection of the parts, thereby evaluating the upper-bound accuracy of action recognition under the proposed representation. In evaluating the accuracy of human action recognition systems, different dataset partitions have been used [6, 8, 13, 28]. A more common reporting of the performance has been based on the use of leave-one-out cross-validation framework [6, 8, 13]. In our work, we use the leave-one-out cross validation technique to train and test our algorithm.

To transform the sequence of body part configurations of a human action to motion descriptor images, we quantize the orientation range $\theta \in (0; 2\pi)$ into O bins and the radius r into R bins. This quantization process also affects the performance of the algorithm. If the quantization is very fine, it results in highly discriminant motion descriptors but the within-class variance also increases and vice-versa if coarse quantization is used. Figure 4 depicts the variation in recognition rates as a function of quantization in orientation and radius. Empirically, we found that three bins for radius and 12 bins for orientation (3×12) is optimal for the KTH dataset, and two bins for radius and 24 bins for orientation (2×24) is optimal for the UCF Sport dataset.

3.2 Action Recognition Performance

The recognition accuracy of our method along with those obtained from previously published methods is presented in Table 1 for the KTH dataset, which is a standard benchmark for human action recognition. The 95.67% recognition accuracy achieved by our method is comparable to the state-of-the-art methods. From the confusion matrices in Figure 5(L), it is

Approach	Year	Accuracy (%)
Schuldt [28]	2004	71.72
Dollar [8]	2005	81.17
Kim [17]	2007	95.33
Laptev [21]	2008	91.80
Liu [22]	2009	93.80
Gilbert [13]	2009	96.70
Bregonzio [6]	2009	93.17
Yao [34]	2010	92.00
Kovashka and Grauman [19]	2010	94.53
Kaaniche and Bremond [15]	2010	94.67
Cao [7]	2010	95.02
Our method		95.67

Table 1: Comparison of recognition rates on the KTH dataset.

evident that most of the actions are accurately recognized, even actions jogging and running with subtle visual differences, which are easily confused by some other methods. Results on the previous version of the UCF Sport dataset containing 150 videos have sometimes been reported with ten actions [19, 31, 34], but a more consistent reporting has been with nine actions. More recently, results have also been reported for other approaches evaluated on the updated UCF Sport dataset with 200 videos [26, 35]. Our results are based on considering nine actions as presented in this updated dataset. Table 2 shows the comparison between our approach and previous approaches on the UCF Sport dataset. To the best of our knowledge, our results, based on a fully automatic pipeline, provide the best recognition accuracy of 88.83%. Figure 5(R) depicts the confusion matrix across the nine actions. These results demonstrate the discriminative power in our representation of human actions and the efficiency and accuracy of our recognition algorithm. Figure 6 shows few examples of automatic part detection for both KTH and UCF Sport datasets. While the automatic part detection is not always accurate, the results indicate the robustness of the proposed representation and its ability to compensate for errors. However, the evaluation of our approach without errors in part detection would provide an upper bound for the performance of our

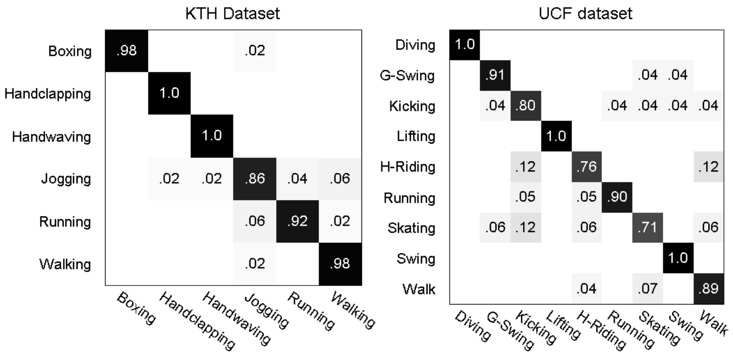

Figure 5: Confusion matrices (L) for the KTH dataset and (R) for the UCF Sport dataset.

Approach	Year	Accuracy (%)
Rodriguez [26]	2008	69.20
Yeffet and Wolf [35]	2009	79.20
Wang [31]	2009	85.60
Yao [34]	2010	86.60
Kovashka and Grauman [19]	2010	87.27
Our method		88.83

Table 2: Comparison of recognition rates on the UCF Sport dataset.

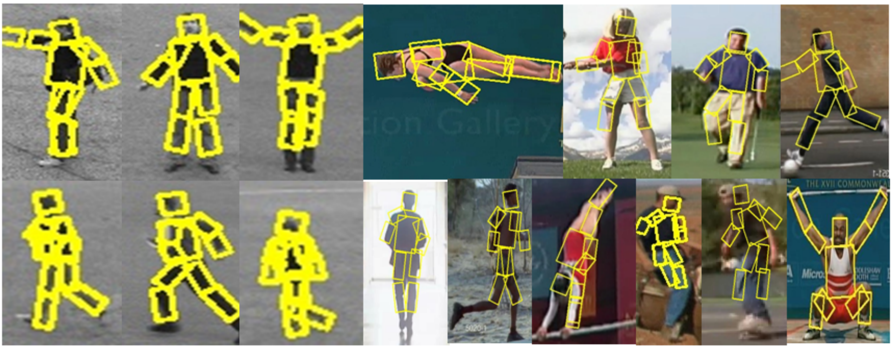

Figure 6: Automatic human body part detection for the KTH and UCF Sport datasets.

approach. Hence, we report the results on both datasets with manually annotation of human body parts. This results in 97.83% recognition accuracy for the KTH dataset and 91.37% for the UCF Sport dataset. The results not only show the upper bound in recognition rate of our method, but also indicate that our approach can achieve better results by improving the human part detection model.

4 Conclusion

In this paper, we have proposed a part-based 2D representation of human action that naturally encodes the temporal progression of an action. We also proposed a robust human action recognition algorithm inspired by sparse representation. Our image representation incorporates both global and local motion information and is robust to variations in poses, view points and human appearance. Experiments have shown the efficiency of our method by obtaining state-of-the-art recognition results on both the benchmark KTH dataset and the challenging UCF Sport dataset.

Acknowledgements

This work was supported in part by the US Department of Justice 2009-MU-MU-K004. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of our sponsors.

References

[1] A. Ali and J.K. Aggarwal. Segmentation and recognition of continuous human activity. In *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, pages 28–35, Vancouver, BC, Canada, 2001.

[2] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1-2):237–260, 1998.

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: people detection and articulated pose estimation. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1014–1021, Miami, FL, Jun. 2009.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. IEEE International Conference on Computer Vision*, pages 1395–1402, Beijing, China, Oct. 2005.

[5] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3): 257–267, Mar. 2001.

[6] M. Bregonzio, S. Gong, and T. Xiang. Recognizing action as clouds of space-time interest points. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1948–1955, Miami, FL, Jun. 2009.

[7] L. Cao, Z. Liu, and T.S. Huang. Cross-dataset action detection. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1998–2005, San Francisco, CA, Jun. 2010.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, Beijing, China, Jan. 2005.

[9] D.L. Donoho. For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006.

[10] S. Eickeler, A. Kosmala, and G. Rigoll. Hidden markov model based continuous online gesture recognition. In *Proc. International Conference on Pattern Recognition*, pages 1206–1208, Washington, DC, USA, 1998.

[11] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[12] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[13] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proc. IEEE International Conference on Computer Vision*, pages 925–931, Kyoto, May 2009.

[14] Y. Iwai, H. Shimizu, and M. Yachida. Real-time context-based gesture recognition using HMM and automaton. In *Proc. IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real -Time Systems*, pages 127 – 134, Corfu , Greece, 1999.

[15] M.B. Kaaniche and F. Bremond. Gesture recognition by learning local motion signatures. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2745–2752, Los Alamitos, CA, USA, 2010.

[16] V. Kellokumpu, M. Pietikainen, and J. Heikkilä. Human activity recognition using sequences of postures. In *Proc. Machine Vision Applications*, pages 570–573, Tsukuba Science City, Japan, 2005.

[17] T. Kim, S. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, MN, Jun. 2007.

[18] T. Kobayashi and N. Otsu. Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation. In *Proc. International Conference on Pattern Recognition*, volume 4, pages 741–744, Washington, DC, USA, 2004.

[19] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2046–2053, San Francisco, CA, Jun. 2010.

[20] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, pages 107–123, 2005.

[21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, Jun. 2008.

[22] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1996 –2003, Miami, FL, Jun. 2009.

[23] T.B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, pages 90 – 126, 2006.

[24] J.C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Los Alamitos, CA, USA, 2007.

[25] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, pages 976 – 990, 2010.

[26] M.D. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1 – 8, Anchorage, AK, Jun. 2008.

[27] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 422 – 429, San Francisco, CA, Jun. 2010.

[28] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proc. International Conference on Pattern Recognition*, pages 32 – 36, Los Alamitos, CA, USA, 2004.

[29] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 405 – 412, San Diego, CA, 2005.

[30] K.N. Tran, I.A. Kakadiaris, and S.K. Shah. Fusion of human postures for continuous action recognition. In *Proc. European Conference on Computer Vision Workshop on Sign, Gesture, and Activity*, Crete, Greece, 2010.

[31] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. British Machine Vision Conference*, pages 127–138, London, UK, 2009.

[32] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pages 210 –227, 2009.

[33] S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3373 –3376, 2008.

[34] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2061 –2068, San Francisco, CA, 2010.

[35] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *Proc. IEEE International Conference on Computer Vision*, pages 492 –497, Kyoto, Japan, May 2009.

[36] M. Ziaeefard and H. Ebrahimnezhad. Hierarchical human action recognition by normalized-polar histogram. In *Proc. International Conference Pattern Recognition*, pages 3720 –3723, Istanbul, Turkey, Aug. 2010.