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

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. Ideally, the desired representation for actions should generalize over variations in viewpoint, human appearance, and spatio-temporal changes. Human action representation can be divided into two categories: global representations and local representations [3]. The global representations can encode much of the information but they are more sensitive to the environment. Local representations are less sensitive to the environment but they depend on the accuracy of interest point detectors. 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 [1] and the framework of sparse representations for recognition [4]. 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. Figure 1 depicts the overview of our approach.

We propose a 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 forms a description of the relative motion of body parts that constitutes a specific action. The motion descriptor of each body part 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 image. Thus, every human action is represented by P motion descriptor images, each describing the motion of the P body parts.

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 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.,

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

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

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.,

$$A^p = [A_1^p, ..., A_K^p] = [a_{1,1}^p, a_{1,2}^p, ..., a_{K,n_k}^p]. \tag{2}$$

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

$$y^p = A^p x_0^p \in \mathbb{R}^m, \qquad p = 1, ..., P \tag{3}$$

where $x_0^p = [0, ..0, w_{k,1}^p, w_{k,2}^p, ..., w_{k,n_k}^p, 0, ..., 0]^T \in \mathbb{R}^n (n = \sum_{k=1}^K n_k)$ is the coefficient vector whose elements are zero except for some elements that are associated with the k^{th} class.

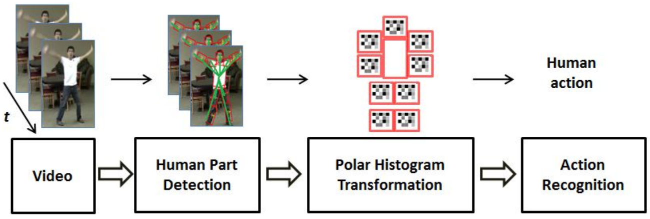

Figure 1: Overview of the human action recognition system.

This representation is naturally sparse if the number of action classes K is reasonably large. The more sparse is x_0^p, the easier it is to accurately determine the action class of test sample y^p for a 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:

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

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

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 [2]

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

Given a new test motion descriptor y^p for particular part p, we first compute its representation $\hat{x}_1^p$ via Eq. 5. For each class k, let us define function $\phi_k^p : \mathbb{R}^n \to \mathbb{R}^n$ 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:

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

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] 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.

[2] 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.

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

[4] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, Feb. 2009.