An Invariant Model of the Significance of Different Body Parts in Recognizing Different Actions

Yuping Shen and Hassan Foroosh

Abstract

In this paper, we show that different body parts do not play equally important roles in recognizing a human action in video data. We investigate to what extent a body part plays a role in recognition of different actions and hence propose a generic method of assigning weights to different body points. The approach is inspired by the strong evidence in the applied perception community that humans perform recognition in a foveated manner, that is they recognize events or objects by only focusing on visually significant aspects. An important contribution of our method is that the computation of the weights assigned to body parts is invariant to viewing directions and camera parameters in the input data. We have performed extensive experiments to validate the proposed approach and demonstrate its significance. In particular, results show that considerable improvement in performance is gained by taking into account the relative importance of different body parts as defined by our approach.

I. INTRODUCTION

Human action recognition from video data has a wide range of applications in areas such as surveillance and image retrieval [10], [13], [14], [58]–[61], [67], [119], [121], [122], [124], image annotation [125]–[129], video post-production and editing [5], [6], [20], [26], [50], [84], and self-localization [62], [63], [68]–[70], to name a few.

The literature on human action recognition from video data includes both monocular and multiple view methods [9], [12], [15], [16], [35], [112]–[117], [120], [123]. Often, multiple view methods are designed to tackle viewpoint invariant recognition [9], [12], [15], [16], [13]–[17], [117], although such methods may require calibration across views [8], [25], [56], [57], [64], [65], [71]–[74], image registration [7], [17]–[19], [21]–[24], [27], [28], [43]–[49], [55], [103], [104], or tracking across views [79]–[81], [107], [118]. There are also methods that rely on human-object interaction [87], [135], [136], which often require identifying image contents other than humans [2], [4], [36]–[38], [40], [42], [76], [78], [85], [134]. Other preprocessing steps that may be needed include image restoration [29]–[31], [53], [77]–[79], [92]–[94], [96]–[102], [105], [106], [108], or scene modeling [11], [12], [34], [66].

In this paper, we look at a very specific problem of determining which body parts play a role in recognizing an action and to what extent. This is a crucial information that may serve as a prior information to any method in the literature, whether video-based or single-image method.

II. RELATED WORK

The literature in human action recognition has been extremely active in the past two decades and significant progress has been made in this area [51], [82], [83], [131]. Human action recognition methods start by assuming a model of the human body, e.g. silhouette, body points, stick model, etc., and build algorithms that use the adopted model to recognize body pose and its motion over time. While there has been many different ways of classifying existing action recognition methods in the literature [51], [82], [83], [131], we present herein a different perspective. We consider action recognition at three different levels: (i) the subject level, which requires studying the kinematics of the articulated human body; (ii) the image level, which requires studying the process of imaging the 3D human body both in terms of the geometry and lighting effects; and (iii) the observer level, which requires studying how an observer would interpret visual data and recognize an action. This way of categorizing action recognition is rather uncommon but offers a different point of view of how one can approach this problem.

Herein we are interested in investigating observer level issues from a new point of view. It has been long argued in the applied perception community [90] that human observers use a foveated strategy, which may be interpreted as an approach using “importance sampling”. What this implies is that humans sample only the most significant aspects of an event or action for recognition, and do not give equal importance to every observed data point. Most of existing action recognition methods in the literature have primarily focused on subject level [33], [34], [86], [132], [133], [137], [138] and image level [41], [75], [89], [130], [134] issues. A number of methods have focused on the observer level [1], [89], but mostly from a machine learning point of view. An interesting work by Sheikh et al. [91] recently has perhaps a close connection to what we aim to address.

Yuping Shen was with the Department of Computer Science, University of Central Florida, Orlando, FL, 32816 USA at the time this project was conducted. (e-mail: ypshen@cs.ucf.edu).

Hassan Foroosh is with the Department of Computer Science, University of Central Florida, Orlando, FL, 32816 USA (e-mail: foroosh@cs.ucf.edu).
in this paper. However, they investigate a holistic approach in an abstract framework of action space. Our goal is to propose an observer level approach with more direct connection to image level features, which we believe would be more “intuitive”. In other words, we are interested in exploring an idea similar to importance sampling where image level features are directly assigned different importance weights that are derived directly from the extent to which the feature affects the recognition performance for a given action.

Fig. 1. An example of human poses, with the points used for body representation: head, shoulders, elbows, hands, knees, and feet. A body point triplet is shown.

One particular issue that has attracted attention among many of these groups is invariance \cite{39, 52, 55, 86, 88, 110, 111}. Invariance is an important issue in any recognition problem from image or video data, since appearance is substantially affected by the imaging level, i.e. the viewing angle, camera parameters, lighting, etc. In action recognition the issue of invariant recognition is compounded with additional complexity introduced by the high degrees of freedom of human body \cite{139}.

Our focus in this paper is twofold: to investigate action recognition at the observer level which takes into account how humans recognize actions by paying attention to only the most significant visual cues, and to introduce this approach within an invariant framework

III. Our Body Model

We use a point-based model similar to \cite{110, 111}, which consists of a set of 11 body points (see Figure 1). We also adopt the idea of decomposing the non-rigid motion of human body into rigid motions of planes given by triplets of body points \cite{110, 111}. This implies that the non-rigid motion of human body is described by a set of homographies induced by planes associated with the body point triplets. As a result the complex problem of estimating the human body motion is reduced to a set of linear problems associated with the motions of planes.

With the triplet representation of human pose and action, we may consider the relative importance of body point triplets for different actions. For instance, the triplets composed of shoulders and hips have similar motion in walking and jogging, and thus make trivial contribution to distinguish them, while other triplets that consist of shoulder, knee and foot joints carry more essential information of the differences between walking and jogging (See Figure 2). Understanding the roles of body point triplets in human motion/action could help us retrieve more accurate information on human motion, and thus improve the performance of our recognition methods.

Our problem can be described as follows. We have a database of reference actions. Given a target sequence we are interested in finding out which reference sequence in the database best matches our target sequence. We are interested in performing this task in a manner that is invariant to viewing directions and the camera parameters. Our first step is to align the target sequences to be recognized to the reference sequences in our database, as described in the next section.

A. Sequence Alignment

The body motion perceived in two different frames is often referred to as body pose transition. Our sequence alignment method is based on finding corresponding pose transitions in two actions sequences viewed by two different cameras. Our starting point is the view-invariant similarity measure that was proposed in \cite{110, 111}: Any triplet of body points in one camera image and the corresponding triplet of points in a second camera define a homography $H_1$ between the two cameras. After a pose transition the triplets move to new positions defining a new homography $H_2$. If the triplet motions are similar, then the two homographies become consistent with the fundamental matrix, and as a result the homography defined by $H = H_1H_2^{-1}$
will become a homology, two of whose eigenvalues are equal. The equality of the two eigenvalues of the homography $H$, provides thus a measure of similarity for the motion of the two point triplets, which is invariant to viewing angles and the camera parameters [110]. Given 11 body points as shown in Figure 1, there are 165 such triplets, all of which are used to provide a measure of similarity for the two actions being compared. We use the following measure of similarity over all possible triplets for evaluating the similarity of two pose transitions $T_1$ and $T_2$:

$$S(T_1, T_2) = \sum_{i} \frac{|a_i - b_i|}{|a_i + b_i|}$$

where $a_i$ and $b_i$ are the two closest eigenvalue of the homography defined by the $i$-th triplet, and $i = 1, ..., 165$ for 11 body points.

Given a target sequence of $m$ pose transitions $A_j$, $j = 1, ..., m$ and a reference sequence of $n$ pose transitions $B_{j'}$, $j' = 1, ..., n$, in order to find the optimal alignment $\psi : A \rightarrow B$ between the two actions $A$ and $B$, we build a matching error matrix using (reffeq:costf2), and use dynamic programming to find the optimal mapping. Once two sequences are aligned, the main issue is to determine how similar the two actions are. In the next section, we describe our weighting-based action recognition method.
IV. WEIGHTING-BASED HUMAN ACTION RECOGNITION

To study the roles of body-point triplets in action recognition, we select two different sequences of walking action $WA = \{I_{1..l}\}$ and $WB = \{J_{1..m}\}$, and a sequence of running action $R = \{K_{1..n}\}$. We then align sequence $WB$ and $R$ to $WA$, using the alignment method described in Section 2, and obtain the corresponding alignment/mapping $\psi : WA \rightarrow WB$ and $\psi' : WA \rightarrow R$. As discussed in Section 2, the similarity of two poses is computed based on error scores of all body-point triplets motion. For each matched poses $(I_i, J_{\psi(i)})$, we stack the error scores of all triplets as a vector $V_e(i)$:

$$V_e(i) = \begin{bmatrix} E(\Delta_1) \\ E(\Delta_2) \\ \vdots \\ E(\Delta_T) \end{bmatrix}, \quad (2)$$

where $E(\Delta_i) = |a_i - b_i| / (a_i + b_i)$.

We then build an error score matrix $M_e$ for alignment $\psi_{WA \rightarrow WB}$:

$$M_e = \begin{bmatrix} V_e(1) & V_e(2) & \cdots & V_e(l) \end{bmatrix}. \quad (3)$$

Each row $i$ of $M_e$ indicates the dissimilarity scores of triplet $i$ across the sequence, and the expected value of each column $j$ of $M_e$ is the dissimilarity score of pose $I_j$ and $J_{\psi_{WA \rightarrow WB}(j)}$. Similarly we build an error score matrix $M'_e$ for alignment $\psi_{WA \rightarrow R}$. $M_e$ and $M'_e$ are illustrated visually in Figure 3.

![Fig. 3. Visual illustration of $M_e$ (left) and $M'_e$ (right)](image_url)

To study the role of a triplet $i$ in distinguishing walking and running, we compare the $i$-th row of $M_e$ and $M'_e$, as plotted in Figure 4(a) - (f). We found that, some triplets such as triplets 1, 21 and 90 have similar error scores in both cases, which means the motion of these triplets are similar in walking and running. On the other hand, triplets 55, 94 and 116 have high error scores in $M_e$ and low error scores in $M'_e$, that is, the motion of these triplets in a running sequence is different from their motion in a walking sequence. Triplets 55, 94 and 116 reflect the variation in actions of walking and running, thus are more informative than triplets 1, 21 and 90 for the task of distinguishing walking and running actions.

In the following experiments, we compare sequences of different individuals performing the same action, and study the roles of triplets in categorizing them in the same group of action: Select four sequences $G0$, $G1$, $G2$, and $G3$ of golf-swing action, and align $G1$, $G2$, and $G3$ to $G0$ using the alignment method described in Section 2, and then build error score matrix $M^1_e$, $M^2_e$, $M^3_e$ correspondingly as in above experiments. From the illustrations of $M^1_e$, $M^2_e$, $M^3_e$ in Figure 5(a), (b) and (c). The dissimilarity scores of some triplets, such as triplet 120 (see Figure 5(f)), is very consistent across individuals. Some other triplets such as triplets 20 (Figure 5(d)) and 162 (Figure 5(e)) have various error score patterns across individuals, that is, these triplets represent the variations of individuals performing the same action.

**Definition 1:** If a triplet reflects the essential differences between an action $A$ and other actions, we call it a significant triplet of action $A$. All triplets other than significant triplets are referred to as trivial triplets of action $A$. 
A typical significant triplet should (1) convey the variations between actions and/or (2) tolerate the individual variations of the same action. For example, triplets 55, 94 and 116 are significant triplets for walking action, and triplet 20 is a significant triplet for the golf-swing action.

Intuitively, in the task of action recognition, we should place more focus on the significant triplets while reducing the negative impact of trivial triplets, that is, assigning appropriate influence factor to the body-point triplets. In our approach to action recognition, this can be achieved by assigning appropriate weights to the similarity errors of body point triplets in equation (1). That is, equation (1) could be rewritten as:

\[
E(I_1 \rightarrow I_2, J_i \rightarrow J_j) = \sum_{\Delta_i} (\lambda_i \cdot E(\Delta_i)),
\]

where \(\lambda_1 + \lambda_2 + \ldots + \lambda_T = 1, T = \binom{n}{3}, n\) is the number of body points in the human body model.

The next question is, how to determine the optimal set of weights \(\lambda_i\) for different actions. Manual assignment of weights could be biased and difficult for a large database of actions, and is inefficient when new actions are added in. Automatic assignment of weight values is desired for a robust and efficient action recognition system. To achieve this goal, we propose to use a fixed size dataset of training sequences to learn weight values. Suppose we are given a training dataset \(T\) which consists of \(K \times J\) action sequences for \(J\) different actions, each of which with \(K\) pre-aligned sequences performed by various individuals. Let \(\lambda_i^j\) be the weight value of body joint with label \(i\) (\(i = 1 \ldots n\)) for the action \(j\) (\(j = 1 \ldots J\)). Our goal is to find optimal assignment of \(\lambda_i^j\) which maximize the similarity error between sequences of different actions and minimize those of same actions. Since the size of the dataset and the alignments of sequences are fixed, this turns out to be an optimization problem on \(\lambda_i^j\). Our task is to define a good objective function \(f(\lambda_1^1, \lambda_2^1, \ldots, \lambda_n^J)\) for this purpose, and to apply optimization to solve the problem.

### A. Weights on Triplets versus Weights on Body Points

Given a human body model of \(n\) points, we could obtain at most \(\binom{n}{3}\) triplets, and need to solve a \(\binom{n}{3}\) dimensional optimization problem for weight assignment. Even with a simplified human body model of 11 points, this yields a extremely high dimensional \((\binom{11}{3} = 165\) dimensions) problem. On the other hand, the body point triplets are not independent of each other. In fact, adjacent triplets are correlated by their common body points, and the importance of a triplet is also determined...
by the importance of its three corners (body points). Therefore, instead of using \( \binom{n}{3} \) variables for weights of \( n \) triplets, we assign \( n \) weights \( \omega_1, \ldots, \omega_n \) to the body points \( P_1, \ldots, P_n \), where:

\[
\omega_1 + \omega_2 + \ldots + \omega_n = 1. \tag{5}
\]

The weight of a triplet \( \Delta = \langle P_i, P_j, P_k \rangle \) are then computed as:

\[
\lambda_{\Delta} = \frac{\omega_i + \omega_j + \omega_k}{\binom{n-1}{2}} = \frac{2(\omega_i + \omega_j + \omega_k)}{(n-1)(n-2)}. \tag{6}
\]

Note that the definition of \( \lambda \) in (6) ensures that \( \lambda_1 + \lambda_2 + \ldots + \lambda_T = 1 \). Using (6), equation (4) is rewritten as:

\[
\bar{S}(A, B) = \frac{2}{(n-1)(n-2)} \sum_{1 \leq i < j < k \leq n} (\omega_i + \omega_j + \omega_k) \cdot E(l, \psi(l))(\Delta_{i,j,k}). \tag{7}
\]

By introducing weights \( \{\omega_1, \ldots, \omega_n\} \) to body points, we reduce the high dimensional optimization problem to a lower dimensional, and more tractable problem.

### B. Automatic Adjustment of Weights

Before moving on to the automatic adjustment of weights, we first discuss the similarity score of two pre-aligned sequences. Given two sequences \( A = \{I_1, \ldots, I_N\} \), \( B = \{J_1, \ldots, J_M\} \), and the known alignment \( \psi : A \rightarrow B \), the similarity of \( A \) and \( B \) is:

\[
\mathcal{S}(A, B) = \sum_{l=1}^{N} S(l, \psi(l)) = N \tau - N \sum_{l=1}^{N} E(I_{l \rightarrow r_1}, J_{\psi(l) \rightarrow r_2}), \tag{8}
\]

where \( r_1 \) and \( r_2 \) are computed reference poses, and \( \tau \) is a threshold, which we set as suggested in \[110\], \[111\]. Therefore, the proximate similarity score of \( A \) and \( B \) is:

\[
\bar{S}(A, B) = N \tau - 2 \cdot N \sum_{1 \leq i < j < k \leq n} (\omega_i + \omega_j + \omega_k) \cdot E(l, \psi(l))(\Delta_{i,j,k}). \tag{9}
\]
Considering that $N$, $\tau$, $n$ and $E^l,\psi(l)\{\Delta_{i,j,k}\}$ are constants given the alignment $\psi$, equation (9) can be further rewritten into a simpler form:

$$\mathcal{F}(A, B) = a_0 - \sum_{i=1}^{n-1} a_i \cdot \omega_i,$$

where $\{a_i\}$ are constants computed from (9).

Now let us return to our problem of automatic weights assignment for action recognition. As discussed earlier, a good objective function would reflect the intuition that, significant triplets should be assigned higher weights, while trivial triplets should be assigned lower weights. Suppose we have a training dataset $\mathcal{T}$ which consists of $K \times J$ action sequences for $J$ different actions, each of which with $K$ pre-aligned sequences performed by various individuals. $\mathcal{T}_k^j$ is the $k$-th sequence in the group of action $j$, and $\mathcal{R}^j$ is the reference sequence of action $j$. To find the optimal weights for action $j$, we define the objective function as:

$$f^j(\omega_1, \omega_2, \ldots, \omega_{n-1}) = Q_1 + \alpha Q_2 - \beta Q_3,$$

where $\alpha$ and $\beta$ are non-negative constants and

$$Q_1 = \frac{1}{K} \sum_{k=1}^{K} \mathcal{F}(\mathcal{R}^j, \mathcal{T}_k^j),$$

$$Q_2 = \frac{1}{K} \sum_{k=1}^{K} \mathcal{F}(\mathcal{R}^j, \mathcal{T}_k^j)^2 - \frac{1}{K^2} \left( \sum_{k=1}^{K} \mathcal{F}(\mathcal{R}^j, \mathcal{T}_k^j) \right)^2,$$

$$Q_3 = \frac{1}{K(J-1)} \sum_{1 \leq i < j, i \neq j} \sum_{k=1}^{K} \mathcal{F}(\mathcal{R}^j, \mathcal{T}_k^j).$$

The optimal weights for action $j$ are then computed using:

$$\langle \omega_1, \omega_2, \ldots, \omega_{n-1} \rangle = \arg\max_{\omega_1, \omega_2, \ldots, \omega_{n-1}} f^j(\omega_1, \omega_2, \ldots, \omega_{n-1}).$$

In this objective function, we use $\mathcal{T}_k^j$ as the reference sequence for action $j$, and the term $Q_1$ and $Q_2$ are the mean and variance of similarity scores between $\mathcal{T}_1^j$ and other sequences in the same action. $Q_3$ is the mean of similarity scores between $\mathcal{T}_1^j$ and all sequences in other different actions. Hence $f^j(\omega_1, \omega_2, \ldots, \omega_{n-1})$ achieves high similarity scores for all sequences of same action $j$, and low similarity scores for sequences of different actions. The second term $Q_2$ may be interpreted as a regularization term to ensure the consistency of sequences in the same group.

As $Q_1$ and $Q_3$ are linear functions, and $Q_2$ is quadratic polynomial, our objective function $f^j(\omega_1, \omega_2, \ldots, \omega_{n-1})$ is quadratic polynomial function, and the optimization problem becomes a quadratic programming (QP) problem. There are a variety of methods for solving the QP problem, including interior point, active set, conjugate gradient, etc. In our problem, we adopted the conjugate gradient method, with the initial weight values set to $\langle \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \rangle$.

V. EXPERIMENTS

In this section, we apply the proposed weighting based approach to the action recognition problem, and compare its performance with non-weighting methods proposed in [109–111]. For comparison, we use the same MoCap testing data as in [109–111], and build a MoCap training dataset which consists of total of $2 \times 17 \times 5 = 170$ sequences for 5 actions (walk, jump, golf swing, run, and climb); each action is performed by 2 subjects, and each instance of action is observed by 17 cameras set up different random locations. The same 11-point human body model in Figure 1 is adopted in the training data. We use the same set of reference sequences for the 5 actions, and align the sequences in the training set against the reference sequences.

To obtain optimal weighting for each action $j$, we first aligned all testing sequences against the reference sequence $\mathcal{R}^j$, and stored the similarity scores of triplets for each pair of matched poses. The objective function $f^j(\omega_1, \omega_2, \ldots, \omega_{10})$ is then built based on equation (14), and the computed similarity scores of triplets in the alignments. $f^j(\cdot)$ is a 10-dimensional function, and the weights $\omega_i$ are constrained by

$$\begin{cases} 0 \leq \omega_i \leq 1, & i = 1 \ldots 10, \\ \sum_{i=1}^{10} \omega_i \leq 1. \end{cases}$$

The optimal weights $\langle \omega_1, \omega_2, \ldots, \omega_{10} \rangle$ are then searched to maximize $f^j(\cdot)$, with the initialization at $\langle \frac{1}{10}, \frac{1}{10}, \ldots, \frac{1}{10} \rangle$. The conjugate gradient method is then applied to solve this optimization problem.

After performing the above steps for all the actions, we obtained a set of weights $\mathcal{W}^j$ for each action $j$ in our database. In order to compare our results with existing unweighted methods, we then carried out full comparisons with two methods
Fig. 6. Examples of computed weights. (1) and (2) are computed weights for walking and jumping correspondingly based on fundamental ratio invariant; (3) and (4) are computed weights for walking and jumping correspondingly based on eigenvalue equality invariant.

TABLE I
CONFUSION MATRIX USING EIGENVALUE EQUALITY INVARIANT: LARGE VALUES ON THE DIAGONAL ENTRIES INDICATE ACCURACY.

| Ground-truth | Recognized as | Walk | Jump | Golf Swing | Run | Climb |
|-------------|--------------|------|------|------------|-----|-------|
| Walk        | 46           | 1    | 1    | 2          | 1   | 1     |
| Jump        | 1            | 48   | 1    | 1          |     |       |
| Golf Swing  | 1            |      | 48   | 1          |     |       |
| Run         | 2            | 1    |      | 48         | 44  |       |
| Climb       | 4            | 1    | 1    | 42         |     |       |

proposed recently in the literature: action recognition using the fundamental ratios [109], and the method proposed in [110], [111]. As these methods behave differently, the objective functions we obtain for estimating weights may contain slightly different sets of coefficients. Figure 6 shows the computed weights for walking and jumping when using for the two methods. Although the weights were slightly different, as shown in the figure similar patterns emerged in terms of significant and trivial triplets: same triplets have relatively high weights in both results.

We repeated all the action recognition experiments reported in [109], [110], [111] using the CMU MoCap data and compared our performance. Results are summarized in Tables I and II using the CMU MoCap data and compared our performance. The overall recognition rate is 93.6% using a weighted eigenvalue method, and 90.4% using a weighted fundamental ratios method, which are improved by 2% and 8.8% compared to the unweighted cases, respectively.

VI. CONCLUSION

We propose a generic method of assigning weights to body points using a small set of training set, in order to improve action recognition from video data. Our method is motivated by the studies in applied perception community on selective processing of visual data for recognition. Our experimental results strongly support our hypothesis that weighting body points differently for different actions leads to significant improvement in performance. Furthermore, since our formulation is based on invariant features, our method shows outstanding performance in the presence of varying camera orientations and parameters. Finally, we believe similar frameworks can be applied to other body models such as silhouette, motion flow, and stick-model.

TABLE II
CONFUSION MATRIX USING FUNDAMENTAL RATIOS INVARIANT: LARGE VALUES ON THE DIAGONAL ENTRIES INDICATE ACCURACY.

| Ground-truth | Recognized as | Walk | Jump | Golf Swing | Run | Climb |
|-------------|--------------|------|------|------------|-----|-------|
| Walk        | 45           | 1    | 1    | 2          | 1   | 1     |
| Jump        | 2            | 47   | 1    | 1          |     |       |
| Golf Swing  | 1            |      | 47   | 1          | 1   |       |
| Run         | 3            | 1    |      | 45         | 1   |       |
| Climb       | 4            | 1    | 1    | 2          | 42  |       |
[126] Amara Tariq and Hassan Foroosh. Feature-independent context estimation for automatic image annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1958–1965, 2015.
[127] Amara Tariq, Asim Karim, and Hassan Foroosh. A context-driven extractive framework for generating realistic image descriptions. IEEE Transactions on Image Processing, 26(2):619–632, 2002.
[128] Amara Tariq, Asim Karim, and Hassan Foroosh. Nelasso: Building named entity relationship networks using sparse structured learning. IEEE Trans. on on Pattern Analysis and Machine Intelligence, 2017.
[129] Amara Tariq, Asim Karim, Fernando Gomez, and Hassan Foroosh. Exploiting topical perceptions over multi-lingual text for hashtag suggestion on twitter. In The Twenty-Sixth International FLAIRS Conference, 2013.
[130] L. Wang. Abnormal Walking Gait Analysis Using Silhouette-Masked Flow Histograms. ICPR'06, 3:473–476, 2006.
[131] L. Wang, W. Hu, T. Tan, et al. Recent developments in human motion analysis. Pattern Recognition, 36(3):585–601, 2003.
[132] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In CVPR’07, pages 1–8, 2007.
[133] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. IEEE Trans. PAMI, 25(12):1505–1518, 2003.
[134] Min Wang, Baoyuan Liu, and Hassan Foroosh. Factorized convolutional neural networks. arXiv preprint arXiv:1608.04337, 2016.
[135] Bangpeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(9):1691–1703, 2012.
[136] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In Computer Vision (ICCV), 2011 s, pages 1331–1338. IEEE, 2011.
[137] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. CVPR’05, 1, 2005.
[138] A. Yilmaz and M. Shah. Matching actions in presence of camera motion. CVIU, 104(2-3):221–231, 2006.
[139] V.M. Zatsiorsky. Kinematics of Human Motion. Human Kinetics, 2002.
[140] G. Zhu, C. Xu, W. Gao, and Q. Huang. Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine. LNCS, 3979:89, 2006.