Human Body Articulation for Action Recognition in Video Sequences

Tuan Hue Thi, Sijun Lu, Jian Zhang
National ICT of Australia & University of New South Wales
223 Anzac Parade, Kensington, 2032, NSW, Australia
{TuanHue.Thi,Sijun.Lu,Jian.Zhang}@nicta.com.au

Li Cheng
Toyota Technological Institute at Chicago
1427 East 60th Street, Chicago, Illinois 60637, USA
licheng@tti-c.org

Li Wang
Southeast University
Nanjing, China
wang.li.seu.nj@gmail.com

Abstract

This paper presents a new technique for action recognition in video using human body part-based approach, combining both local feature description of each body part, and global graphical model structure of the human action. The human body is divided into elementary points from which a Decomposable Triangulated Graph will be built. The temporal variation of human activity is encoded in the velocity distribution of each node in the graph, while the graph structure shows the spatial configuration of all the nodes in the action. Tracking trajectories of unlabeled good feature points are correctly labeled using Maximum a Posteriori probability. Dynamic Programming is then implemented to boost up the exhaustive search for the optimal labeling of unknown body parts and the best possible action. A simple and efficient technique for building the optimal structure of the human action graph is also implemented. Experimental results on the KTH dataset proves the success and potential applications of this proposed technique.

1. Introduction

Human action recognition in video sequences has become undoubtedly popular in the field of computer vision, not only because of its vast applications in different areas of interactive entertainment, assisted care, surveillance and sports annotation, but also of its interesting technical challenges. A typical human action recognition framework normally contains one module for feature representation and another part for classifying the action characteristics. For video sequences recorded from a single camera, most of the works fall into two main trends, while one focuses on the whole contour of the human body (holistic approaches [1, 3, 2]), the other focuses on articulated body parts (body segments [9, 7] and local feature points [5, 11]).

In order to recognize human movement, Bobick and Davis in [1] introduces a motion representation called Motion History Image (MHI). Different motion functions of different points in image sequence are calculated and used to determine the presence of the motion as well as to judge the recency of the motion that present. The temporal templates of each motion will be matched with stored template instances to determine the label of human action.

Using the same concept of MHI, Calderara et al. in [2] define the orientations of gradient calculated from MHI as action signature. It will then be used to represent the object motion variation in time. They then use a mixture of von Mises distributions to align different sequences and do the classification against the defined actions.

Another popular reported work of holistic approach can be seen in Ke et al. [3] where they represent the action in video as volumetric features, which is basically the flow of foreground shape segmented from video. Correlation distance is then used to detect the presence of a predefined action in the whole sequence.

The advantage of all these works lie in the fact that human body contours can be easily obtained from a good foreground segmentation system. However, these methods might not be useful in distinguishing between different viewpoints of the same action, nor they can tell the core characteristics of the human action. These concerns are investigated in another perspective by the discriminative part-based research stream.

Ramanan et al. [9] uses their discriminative appearance limb model to detect and localize human limbs in images. An action is defined by them as a sequence of different
poses, or different combinations of limbs. The motion graph classification is then carried out using pose state-space Hidden Markov Models (HMM).

Instead of using body segments, other groups are interested more in detecting and tracking highly representative points in human body. Laptev et al. in different works [5] have proved the effectiveness of using local space-time features. The action recognition is carried out with a multi-channel non-linear Support Vector Machine (SVM) classification on analyzed space-time pyramids.

Another interesting approach using local feature point tracking for human action recognition is presented by Song et al. in [11] in which they learn an approximate probabilistic model of joint features position and velocity. Similar to [7], they use MAP optimization to determine the pose and motion of the body.

In our work, we turn our attention specially to the problem of recognizing human actions using local feature points tracking similar to the works in [6, 5, 11, 12]. An human action in our definition is basically a temporal configuration of different human poses which, in turns, consist of spatial combination of core body parts. Instead of using too many body joints like in [11], it is more realistic to pick the maximum number of 13 core points on the human body to build the motion model. Figure [1] shows a complete articulation of the two dimensional frontal human body image, consists of Head(H), LeftWrist(LW), RightWrist(RW), LeftHip(LH), RightHip(RH), LeftElbow(LE), LeftWrist(LW), RightWrist(RW), RightElbow(RE), RightKnee(RK), LeftKnee(LK), RightFoot(RF), and LeftFoot(LF).

Using local Kanade-Lucas-Tomasi feature detector [10] together with iterative optical flow grouping [8] to track distinctive points over time, we focus on three main tasks, building the most representative model for each action, solving the human part labeling problem, and classifying human motion into different trained action models. The main contribution of this work is the design of a framework consisting of an algorithm for finding the optimal graphical structure of any human action, a human part labeling module using Dynamic Programming and Belief Propagation on graph, and lastly a multi-class action classifier. The proposed approach deals with human action recognition by combining both local feature description of core parts and global structure of the whole human body.

2. Model of Human Action

An Human Action is considered as a temporal variation of spatial configuration of body element parts. For each human action, we assign a graph \( G = (V, E) \) to represent the spatial characteristics of the body nodes \( V \) and their mutual relationships, implied by the graph edges \( E \). Each graph is described by a joint probability function from its nodes, and can be factored into smaller product groups based on the conditional dependency of its edges. Using the same approach as described by Song et al. in [11], we use decomposable triangulated graph as the graphical model to represent human action. One good reason for this choice is the structure nature of a human body is quite similar to a triangulated graph with only few edges added, secondly, this common type of graph takes into account more interaction dependencies among body parts compared to the original tree-like physical structure. A Decomposable Triangulated Graph (DTG) is composed of different cliques of 3 nodes, where there exists a particular Elimination Order (EO), by which the graph can be reduced into a smaller DTG if one point in EO is reduced, until there is only one clique left.

Figure [1] is a DTG version of the graph in Fig-
one Elimination Order of this graph is $EO = ADFBEGCHIJKL$.

Based on this DTG and its EO, the joint probability function can then be defined as

$$
P_G = \frac{1}{Z} P(A|BC)P(D|BF)P(F|BH)\ P(B|CH)P(E|CG)P(G|CI)P(C|HI)\ P(H|IJ)P(I|JK)P(J|KL)P(KLM)$$

(1)

where $Z$ is added to ensure $P_G$ is a legitimate probability function. Equation (1) shows the spatial structure of all body parts (or human pose), the temporal variation is taken into account by assigning a Gaussian distribution function for each clique of the body graph.

$$
C_i \propto N_i(\mu_i, \Sigma_i)
$$

(2)

$N_i$ is in fact a multivariate Gaussian distribution (with full covariance matrix) built on a feature vector defined by

$$
F = [c \ v \ d]^T
$$

(3)

where $c$ represents the 2D coordinates, $v$ is the velocity, and $d$ is the distances between the nodes, the joint of these 3 features is used to compute the probability of each clique. Bayes Rule is used for conditional probability functions

$$
P(A|BC) = \frac{P(ABC)}{P(BC)}
$$

(4)

### 3. Human Part Labeling

Running KLT Detector together with Optical Flow Tracking on both the training and testing sequences give different 2D trajectory of the representative points available in the sequences (Figure [3]). Region of interests containing the feature points is then normalized as specified in Figure [4], basically, a rectangle with pre-defined size $L \times 2L$ will be created containing all human points with the top point (Head) always in the coordinate $(L, L)$. Normalization at this stage is quite important since absolute and relative positions of these feature points will be used to build the discriminative geometric model. In training stage, those trajectories come with the label of $V_i \subset V$ from which the model parameters $\mu_i, \Sigma_i$ can be calculated.

$$
V = \{H, RW, RE, RS, LW, LE, LS, \ RHF, LH, RK, LF, RF, LF\}
$$

(5)

On the other hand, testing sequences $X = x_t$ does not give any information of which labels they belong to. The idea of human part labeling now becomes finding a matching set $Y = \{ (i_t, x_t) \}$. Dynamic Programming is implemented in sequential stages along the path of the EO to search for local best match for each node, then the global value will be used in the end as the judgement to pick the optimal arrangement of observation in graph model. Define $\Pi_j$ as one possible matching of the observation $X$ into known labels $L$, the objective is to look for the best combination of $\Pi^*$ of $X$ where

$$
\Pi^* = \arg \max_{\Pi_j} P_G(X_{\Pi_j})
$$

(6)

### 4. Event Classification

Based on equation [6], the idea of event classification now becomes comparing the Maximum a Posterior (MAP) of human part motion observation given different action models. Define $\alpha$ as the index of each given human action, the likelihood of each action $A_{\alpha}$ given its best possible labeling of motion observation $X_{\Pi^*_\alpha}$ is computed and normalized to recognize the action being carried out.

$$
L(A_{\alpha}|X_{\Pi^*_\alpha}) = P(X_{\Pi^*_\alpha}|A_{\alpha})
$$

(7)

$$
A_X = \arg \max_{\alpha} L(A_{\alpha}|X_{\Pi^*_\alpha})
$$

(8)
5. Graphical Models for Human Actions

The initial intuitive choice of the action graph structure $G_{A_0}$ is not always the most representative model for the action itself. In fact, depending on the nature of human motion variation, in some actions, some parts of the human might have more contribution to the overall action, while in others, some parts should be directly connected even physically, they are not. Song et al. in [12] introduces an unsupervised algorithm to find the suitable structure for different actions and human poses. While the idea of auto-generation of graph structure in their work is interesting, their technique always assumes the co-existence of all possible nodes in the structure, hence, leads to the ignorance of different importance weights from each node contribution in the overall action. We propose a much simpler but effective way to search for the optimal structure using an iterative maximization technique inspired by the idea of Expectation Maximization (EM) algorithm. As mentioned in Section [2], for each DTG of the action, there exists at least one EO from which the graph can be constructed or rearranged. From each initial action graph, we pick a bottom-up EO for consideration. The algorithm is initialized with the full DTG of the human segments visible in that action, and iteratively for each graph, the Maximum Likelihood (ML) $L(\hat{A}_G|X)$ is computed, then one node is eliminated according to the chosen EO, the new graph is generated from restructuring the previous graph with new node set (Figure [5]).

![Figure 5. Iterative graph structuring along the chosen EO](image)

We also employ the neighboring constraints in the restructuring step, based on the geometric dependency characteristics of the human body parts, we define a fundamental structure of the body as a 2D binary matrix $S_{5 \times 7}$ with horizontal and vertical distances among nodes, a maximum gap threshold of 4 is chosen to decide which nodes could possibly be linked together.

$$S = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$ (9)

Table [1] summarizes the steps taken to derive the most representative graph structure of each human action.

6. Experimental results

The approach presented in this paper is evaluated using the KTH dataset [4] containing 25 different persons performing 6 elementary actions (HandWaving(HW), HandClapping(HC), Boxing(BX), Walking(WK), Jogging(JG), and Running(RN)) under 4 different recording contexts. Since we are more interested in the problem of multi-class classification of different human body parts and different human activities, only a representative portion from each sequence is extracted and used as the sequence inputs for our testing system. We validate our proposed approach by first finding the optimal graph structure representing each kind of human action, then testing the performance of the human labeling and action classification tasks.

6.1. Optimal Human Action Graph Structure

Optimal graphs for 6 actions from KTH dataset have been obtained according to the steps in Table [1]. Interestingly, the optimal graphs for Walking, Jogging and Running actions are the same, reflecting their kinematic similarity.
6.2. Human Part Labeling

The results of part labeling tasks are shown on the last row of Table [3] in terms of event-based order. Few samples of True and False Positive are displayed in Figure [5].

<table>
<thead>
<tr>
<th>(%)</th>
<th>HW</th>
<th>HC</th>
<th>BX</th>
<th>WK</th>
<th>JG</th>
<th>RN</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW</td>
<td>95</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HC</td>
<td>4</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>BX</td>
<td>1</td>
<td>0</td>
<td>85</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>WK</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>85</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>JG</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>75</td>
<td>18</td>
</tr>
<tr>
<td>RN</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>13</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td>86</td>
<td>76</td>
<td>95</td>
<td>89</td>
<td>83</td>
<td>63</td>
</tr>
</tbody>
</table>

Table 3. Human Action Recognition and Part Labeling Results

It can be roughly seen that in simple graph where only few parts are visible (e.g. Boxing with 95%) the accuracy is quite high, meanwhile the event with complicated action and more parts considered, the accuracy can be improved more (e.g. Running with 63%).

6.3. Action Recognition

Event Classification among 6 different actions has been carried out under different configurations of the model to see how the factors within the models affect on the whole classification accuracy. It is shown that normalization constants for each event, as in Equation [1] is really important to avoid the bias from actions with large graph (e.g. Hand-Waving, HandClapping). Also, feature function with velocity included shows better results for events with movement speed variation (e.g. Walking, Jogging, and Running), while in other events, geometric characteristics are quite sufficient to yield satisfactory results. The results of the action recognition task are shown in the Classification Confusion Matrix of Table [3], few samples of True and False Positives are shown in Figure [5].

The results for both part labeling and action recognition tasks in Table [3] show that for actions like HandWaving, HandClapping, and Running, the task of recognizing action gives more accuracy than the labeling itself, meaning giving some false detections of separate parts, the overall action can still be recognized correctly given the correlating motion and dependency between the parts. On the other hand, in actions like Boxing, Walking, and Jogging, the result of action recognition relies on the detection rate in each different parts, which explains why the action recognition task is less accurate than the human part labeling.

7. Conclusions

A new method for classifying different human actions in video has been presented. Tackling action classification and recognition using the atomic elements of each action, human part movement, has been proven to give desirable results. In addition, by doing the classification task, there can be obtained the information of each visible body part in the action. Even with simple feature descriptors (coordinates and velocity in this case), successful results can already be achieved. Further investigations on richer features as well as complex classification techniques, or different graphical models might be investigated to make the framework even more robust.

Acknowledgement

NICTA is funded by Australian Governments Department of Communications, Information Technology, the Arts and the Australian Research Council through Backing Australia’s Ability and the ICT Centre of Excellence program.

References

Table 4. Samples of true and false positives in Human Part Labeling task

Table 5. Samples of true and false positives in Human Action Recognition task


