H-048

動作を表す言葉の語義に従った観測動作の分解に基づく日常生活 での人物動作認識
Interpretation of Human Action in Daily Life Scene based on Action Decomposition using Terms in Dictionary

ロクマン ジュアンダ† 今井 順一† 金子 正秀†
Juanda Lokman Junichi Imai Masahide Kaneko

1. Introduction

As the one of the most active research area in computer vision, human motion analysis has been applied to a wide spectrum of applications, such as, virtual reality, automatic surveillance, smart video retrieval, human and computer/robot interaction, and soon.

Regardless of the applications, human motion analysis system generally can be divided into several subsystems according to their functionalities: detection, tracking and recognition. Detection usually comes as the first processing in order to segment the motion that belongs or contains the human of interest; usually it is composed of motion segmentation (motion and non-motion areas) and motion classification (human and non-human segments). Tracking usually come next after the detection in order to keep tracking the segmented objects in the whole input sequence. After tracking the interested objects (humans), the motion of the interested objects usually are analyzed for just motion classification, action recognition, giving some responses according to the motion (gesture recognition), giving a semantic description or generally behaviors understanding.

The objective of this paper is to understand the human motion in daily life scene using semantic description to describe what is/are going on the video imagery from fined-level description (the meaning of the motion of each body part) to gross-level (the whole activity that happened in the scene) description.

This approach uses the words in the dictionary (in our case we are going to assign a semantic description to the motion of the human, thus we just concentrate on the action verb category) in order to build the word linkage. The words in the dictionary usually are described (explained) using a sentence or phrases. And basically the sentence consists of a verb and one or more noun phrases, each associated with the verb. And the verb in the description (explanation) can be used again to find the meaning of that term. Thus we can build the linkage of one word with the other words that are used to describe/explain the meaning of that word while the noun or the noun phrases is used as the conditions/constraints. This link is continuing until it achieves the primitive (atomic) word that can be mapped directly with the feature extracted/measured from the video imagery. Bayesian Network (BN) is used to represent the linkage of the words. Thus we can assign the semantic description to the analyzed motion through the inference of the BN with the measurement (observed features) from the input sequence as the input evidence to the BN.

In order to show the effectiveness of the proposed framework, we will discuss the differences among our framework and previous related works in section 2. Section 3 will explain the system overview and several assumptions in our system then section 4 presents the action model for mapping the low level features from sequence of images to the high level of semantic description of the scene. The usefulness, effectiveness and robustness of our proposed framework will be proved with some experiments as shown in section 5. Conclusion and possibility of future works will be following in section 6.

† The University of Electro-Communications

2. Related Works

Most of the previous works just concentrated on the motion of whole body (or just a single body part such as hand) without considering any motion of each separate body parts [1,2,3], while our system concentrates on the motion of each body parts. Concentrating on the motion of the whole body just can be used to recognize several simple actions such as walking, running, skipping, etc. The motion of each body parts and relative motion between the body parts can be used to recognize more unstructured (or complex) motions of the body part, and simultaneously to recognize motions for each body parts.

Appearance-based recognition [4,5] has been used for action recognition and has achieved some level of accuracy, while the effectiveness of the methods is quite satisfying but those methods are limited to some point of view (view dependent/variant) such as facing the camera, parallel to the image plane, etc. Our proposed system uses stereo camera to capture the 3D position of the body parts (head and hands), thus it can achieve view invariant, though we are not going to handle the occlusion problem.

Most of the previous works have been done with quite strict constraints (limited environment) for example; office scene, class scene or sport scene (ballet, tennis, soccer, etc) and little concentration on more relax constraint of human motion in daily life environment [7]. Actions in a specific environment is subjected to motions that are limited to the specific scene and which they have the specific terms (terminology) to describe/name the motion, such as in ballet, it has hundreds of term to describe the structured motion in ballet. But human motions in daily life are unstructured and without any terminology.

From the simple point of view, action recognition can be viewed as a classification of changing pattern in time sequence or mapping between the geometric feature extracted from the image sequence and the high-level semantic description. Pattern classification method seems to work well for the case where the motion is structured and slightly variant of the same motion class. This method cannot work well for the motions in the daily life

which are unstructured and very varying. Hierarchical method has been observed in order to map between the geometric feature and high-level semantic description. But the concepts of the hierarchical method are still limited to some degree of accuracy because of the features selection that are used as the constraints based on the appearance, limited poses etc.

3. System Overview

We propose a novel framework that can work well from understanding the action of each body parts (fined-level), simultaneous actions of the human, to understanding the whole activity (gross-level). This method is view invariant because we choose the features that are invariant of viewing (we use a stereo camera).

In this research we concentrate on the motion or action recognition in daily life scene and assign one or more semantic description about the action of the human in the scene, thus we skip the detection and tracking step by using glove for each hand (See Fig. 1).

Figure 1. Sample of Input Image

We believe that human motion in daily life can be recognized just using the hands and head (face) information. The detail of the joint or location of the elbow is not so important (the limb is just used as actuator, e.g., flexor and extensor), but the relative distance (motion) of the hands and hand with the face are much important to discriminate the motion in daily life.

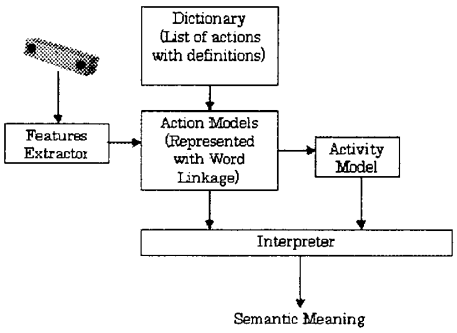


Figure 2. System Overview

Figure 2 shows the overall of our proposed system. Images are inputted from a stereo camera to extract the 3D trajectory of each hand and head (face). Dictionary (in an appropriate representation for computation) is used to construct the action model and activity model, and finally the Interpreter will output the semantic description about recognized action or simultaneous actions of the human using the inference result from Bayesian Network that models the action.

4. Action Model

As mentioned above that usually in human action recognition, the extracted features were used directly to train the system in order to construct the action model. These methods work well with some constraints or conditions only (specific viewpoint, slightly varying in motion, predefined or constraint motion, etc).

In order to overcome the limitation of previous methods, we have to find a competence/suitable mapping between the extracted features (extracted motion cues) from image (or image sequence) and the action to be recognized.

Human motions in daily life are lack of terminology compared to the motions in some sports such as ballet, tennis, etc. For example in ballet, there is a term to describe every single motion (sequence of motions or poses). Fortunately, we have the dictionary that contains the definition of a word in a general manner, and we want to describe the human action using the word that related to the action in dictionary (usually categorized as verb). Thus a more general and robust method to recognize the action(s) and activity in the daily life scene is obtained by using the definition from the dictionary as the hypothesis. The problem is how to map between the extracted cues (features) from the scene (or images sequence) and the words in the dictionary.

4.1 Dictionary Terms

In English grammar, the words can be classified into eight parts of speech: verb, noun, pronoun, adjective, adverb, preposition, conjunction and interjection. In this case of human motion analysis, we just concentrate on verbs because verbs usually relate to some action and indirectly relate to motion (though some actions may not involve any motions).

The idea is using the words list in the dictionary (see Table 1), looking for the words connection (see Table 2) and finding the atomic word that can be used standalone (independent) without the need of explanation from other words (see Table 3). And this atomic word is directly mapped with the extracted features or cues from image (or images sequence). The relationship is shown as the Figure 3.

Table 1. List of main words with the definition from dictionary

Word	Description/Definition from Dictionary
Single person action (without object)	
Sitting (down)	One's weight is supported by one's buttock
Standing (up)	Upright position, supported by one's feet
Waving	Move (one's hand or arm) to and fro
Moving	Change the position or place
Single person action (with object)	
Grasping	Hold firmly
Picking up	Grasp and lift
Carrying	Move while supporting
Two or more persons	
Shaking Hands	Clasp another's hand

Now two main problems in constructing the word linkage: first, connection between one word and other word and second is

map between the extracted features (cues) from image (or image sequence) to the (atomic/primitive) word.

Table 2. List of derived words with definition from dictionary

Word	Description/Definition from Dictionary
Change	Become different
Lift	Raise or be raised to a higher position
Claps	Grasp tightly with one's hand
Raise	Lift or move to a higher position or level

Table 3. Predefined words that directly related to the observed feature from input image.

Word	Description/Definition from Dictionary
Become/be	Enter a certain of state (?)
Hold	Have in one's hand ($hand_{position} \approx object_{position}$)
Support	Bear all or part of the weight of (?)
Upright	Vertical/unbent *(head, center mass & foot in a vertical line)*
Back and forth	Moving from one place to another and back again *(periodic)*
To and fro	Back and forth *(periodic)*
Firmly	In a stable manner *(no change for > T)*
Tightly	Firmly *(no change for > T)*
Vertically	Vertical direction ($Y_t \neq Y_{t-1}$)
Different	Not the same ($x_i \neq x_{i-1}$)

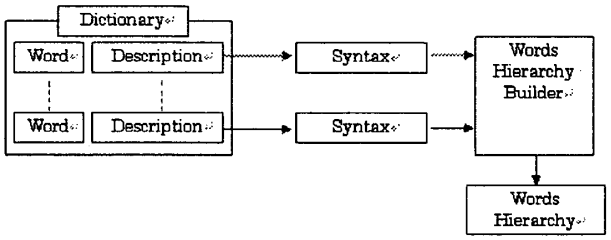


Figure 3. Words linkage from definition that comes from the dictionary

4.2 Expression

In order to build the words linkage automatically and later use for computation, it should be represented in a form that can be easily processed by computer, thus we need to analyze the sentence or phrase of the word definition from the dictionary and transform it into a formal pattern.

One verb can be defined with one or several verbs and some noun phrase includes noun (object) that is manipulated by the action and some required conditions. These conditions could be noun, adjective, adverb and phrase in which the phrase itself can be derived into another verb, etc.

In the part of speech, the same word can be a verb in one sentence, and can be a noun or even adjective in other sentence or context. Thus we need to analyze and decompose the sentence using the method that classifies each noun phrase that accompanying the verb by its semantic role (thematic relations). Expression that we use here is almost same as the case grammar proposed by Fillmore [6], but in our case, we use agent, object, source, goal, and condition(s).

Expression of sentence in the description of one word is shown in a compact form as follow:

$$VERB \triangleq \left\langle {}^{[source]}_{[agent]}\{verb\}^{[goal]}_{[object]} \Big|^{[condition(s)]} \right\rangle \quad (1)$$

Where agent is the body part that does the action (verb), object is something or body part of somebody, source and goal are the source and goal location before and after the action, condition is the requirements or states in order to accomplish the action.

Because this research is just concentrate on action interpretation, how the sentence from definition in dictionary is transformed into expression above is not our main concern, we just do the manually transform the definition from dictionary into that representation. The process of transformation may be concerned as part of natural language processing area.

Table 4 shows the expression for the words in the main words list. Derived words (see Table 5) can be expressed into the same format until the derived words is small enough to be directly mapped to the measured feature from the visual image.

Table 4. Expression for main words

Terms	Expression
Sitting (down)	${}_{one's\ buttock}\{support\}_{weight}$
Standing (up)	${}_{one's\ feet}\{support\}_{weight} \Big|^{upright\ position}$
Waving	$\{move\}_{hand\ /\ arm} \Big|^{to\ and\ fro}$
Moving	${}^{position\ 1}\{change\}^{position\ 2}$
Grasping	$\{hold\}\Big|^{firmly}$
Pick up	$\{grasp\} \wedge \{lift\}$
Carrying	$\{move\} \wedge \{support\}$
Shaking hands	${}_{agent's\ hand}\{clasp\}_{other's\ hand}$

Table 5. Expression for derived words

Terms	Expression
Change	$\{become\}\Big|^{different}$
Lift	$\{raise\}$
Clasp	$\{grasp\}\Big|^{tightly}$
Raise	${}^{low(position)}\{move\}^{high(position)}$

4.3 Word Linkage & Representation

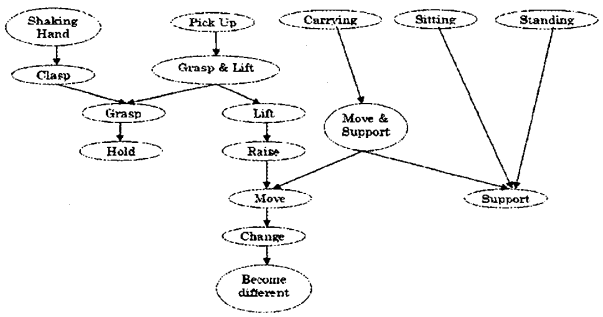


Figure 5. Word linkage for all actions

The idea of building word linkage can be used for reasoning/recognizing the action from the image or sequence of image. The problem is how to represent the linkage into a more formal language that can be implemented later into computer.

We are going to show "shaking hand", "picking up", "carrying", "sitting down", and "standing up" actions for example and show the formal representation of the word linkage.

4.4 Observations (Feature Measurements)

We uses one stereo camera (Bumblebee) to capture the 3D coordinates of each body part (both hands and face), then we have instance position, trajectory, the relative distance and relative motion among them.

Through these information we can map the primitive action/word or other constraints such as (speed, direction, periodicity, etc) with the geometric features (see Table 3 as shown in italic brackets).

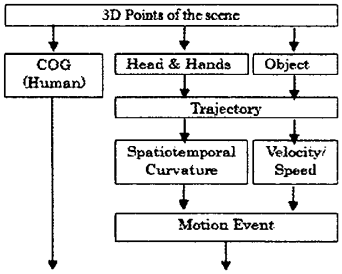


Figure 6. Calculation of geometric features

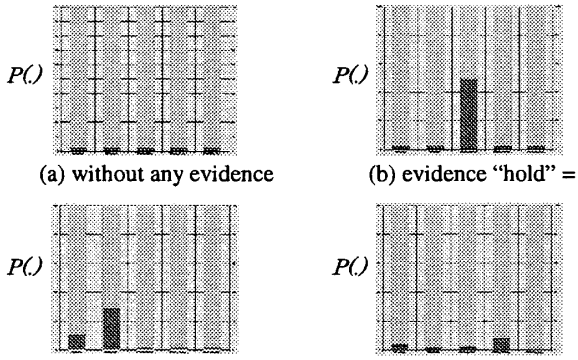
Figure 6 shows how the 3D points of the scene that are calculated from stereo camera are used to calculate the position of the hands, head and object, including the center of gravity (COG of human). From the position of each frame, we can derive the trajectory, spatiotemporal curvature, relative speed, direction, and finally we can derive the motion event from all of these features. All of these features including the motion event are used to map the observed geometric features to the predefined primitive actions/words and constraints words (e.g., frequently, etc.).

5. Simulation Experiment

To show the effectiveness of the proposed system, we use Microsoft Belief Network (MsBNX) to simulate the word linkage representation with Bayesian Network. We manually build the network and assign the conditional probability using some knowledge domain of human actions.

In the simulation, we assume that we can get the measured parameter from the input image. In order to simplify the network, we simply hide the node "change" and "become different". Most of variable of the network just contain 2 states : "yes" and "no", except variable "move" has 4 states : "move higher", "move lower", "move horizontal", and "idle", and variable "support" has 3 states : "support by hand", "support by buttock", and "support by legs".

Figure 7 shows the simulation result for normal condition (without observation/evidences) and with several evidences. The vertical bar shows the probability of the recognized action with scale 0.0 to 1.0 (from left to right): "Carrying", "Picking Up", "Shaking Hand", "Sitting Down", and "Standing Up". Fig. 7(a) shows the result without any evidence, (b) with evidence "hold" = 1, (c) with evidence "move higher", it shows the possibility of "picking up" and "carrying", and (d) with evidence "move horizontal" and "support by buttock".



(a) without any evidence (b) evidence "hold" = 1

(c) evidence "move higher" (d) evidence "move horizontal" and "support by buttock"

Figure 7. Simulation results. (Five bars correspond to actions "carrying", "picking up", "shaking hand", "sitting down", and "standing up"

6. Conclusion

We proposed a novel framework using the definition of human action related words from dictionary in order to find a general, robust and view invariant feature for action recognition. The human actions are decomposed into primitive action using the definition from the dictionary and building the word linkage by representing into Bayesian network. We showed the effectiveness of the proposed method by simulation, how it can recognize human action of each human body part, missing observed data, and how it handles the simultaneous actions of separate body parts.

In order to make an automatic system for recognition we will analyze and map the geometric features from input images into primitive actions.

Reference

[1] Claudette Cedras and Mubarak Shah, Motion-based Recognition: A Survey, Image and Vision Computing, Vol. 13, No. 2, pp. 129-155, Mar 1995.

[2] Thomas B Moeslund and Erik Granum, A Survey of Computer Vision-based Human Motion Capture, Computer Vision and Image Understanding, Vol. 81, No. 3, pp. 231-68, 2001.

[3] Weiming Hu, Tienniu Tan, A Survey on Visual Surveillance of Object Motion and Behaviors, IEEE Trans. On Systems, Man and Cybernetics, Vol. 34, No. 3, pp. 334-352, 2004.

[4] Cen Rao and Mubarak Shah, A View-Invariant Representation of Human Action, Workshop on Detection and Recognition of Events in Video, pp. 55-64, 2001.

[5] A. Kojima, M. Izumi, T. Tamura and K. Fukunaga, Generating Natural Language Description of Human Behavior from Video Images, ICPR, Vol. 4, pp. 728-731, Dec 2000.

[6] Fillmore, C.J., The case for case. In Universals in Linguistic Theory, E. Bach and R. Harms (Eds.). Rinehart and Wiston: New York, pp. 1-88, 1968.

[7] Petkovic, M., Jonker, W. and Zivokic, Z., Recognizing Stroke in Tennis Videos Using Hidden Markov Models, Visual Imaging and Image Processing, 2001.