Human Action Recognition by Inference of Stochastic Regular Grammars

Kyungeun Cho, Hyungje Cho, and Kyhyun Um

Department of Computer and Multimedia Engineering
Dongguk University, Pildong 3 ga 26, Chunggu, Seoul, 100-715, Korea
{cke,chohj,khum}@dgu.ac.kr

Abstract. In this paper, we present a new method of recognizing human actions by inference of stochastic grammars for the purpose of automatic analysis of nonverbal actions of human beings. We applied the principle that a human action can be defined as a combination of multiple articulation movements. We measure and quantize each articulation movements in 3D and represent two sets of 4-connected chain code for xy and zy projection planes, so that they are appropriate for the stochastic grammar inference method. This recognition method is tested by using 900 actions of human upper body. The result shows a comparatively successful achievement of 93.8% recognition rate through the experiments of 8 action types of head and 84.9% recognition rate of 60 action types of upper body.

1 Introduction

Human action recognition is an active area of research in pattern recognition. Medical analysis of human gait movement, nonverbal communication in social psychology, VR using avatar control, automatic man-machine interaction, development of surveillance systems, sign-language recognizer, choreographic analysis of dance and ballet, and gymnastic movement - all belong to this application area of automatic human action recognition. In some areas, action recognition systems are already established such as the Chinese Sensei system analyzing Tai-Chai recognizing and translating sign language [1,2].

Practically, there are manifold phases in recognizing human actions in videos including tracking of human, separation of human bodies from the background, identification of body parts, and recognition of human actions [2]. There are various approaches to the recognition of human actions such as Dynamic Time Warping (DTW), template matching method, fuzzy method [2], Hidden Markov Model (HMM) and syntactic method [3,4] etc.

In a previous research, Stochastic Context Free Grammar as one of the syntactic method was used to the recognition system of human actions [3]. The system consists of an HMM bank and a probabilistic Earley-based parser. Grammar inference was referred to as the further study in their framework of stochastic parsing. In this paper, we show that a stochastic regular grammar inference method can resolve this problem because it has no much limit in inferring the grammar.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 388–396, 2004.

This paper presents a recognition scheme to analyze human actions on 3D temporal data of video where an unsupervised inference procedure is introduced to stochastic grammars. This scheme is based on the principle that a human action is defined as a combination of multiple articulation movements which is built up from multiple mutually-synchronized temporal data [5]. Human action is considered as a stochastically predictable sequence of states [3]. So, we apply a mixed statistical-syntactic approach to the recognition of human upper body action. In addition, we use a mechanism to infer stochastic grammars, which deals with learning the production probabilities for a set of given grammars. A grammar represents single human action.

The remainder of this paper is organized as follows: Section 2 recalls the theoretical concepts and notations of stochastic grammar inference method. Section 3 presents a data representation for human action sequence used in our work. Section 4 explains an overview of action recognizer and describes its main functions. Section 5 shows experimental results and analysis. Finally, in the last section we provide some discussion about our contribution and its future works.

2 Inference of Stochastic Grammar

In our work, we apply an inference of stochastic grammar scheme to recognize human actions. In this section we concisely describe the appropriateness of our method for human action recognition and the theoretical concepts and notations of stochastic grammar inference method.

2.1 Syntactic Pattern Recognition for Human Actions

Syntactic pattern recognition is a method focused on the structure of pattern. The recognition method dismantles an objective pattern into simple subpatterns to recognize and explain relations among the subpatterns with a grammar theory of formal language. In general a movement of one body part in a human action can be described as a sequence of subpatterns, such as left-left-left-right-right-right-up-down. In this case left, right, up, down are subpatterns that construct a movement of one body part. If a subpattern sequence includes transformations, noises, observatory errors, and incomplete feature extractions, etc., it is very difficult to express all actions in simple grammar although they have the same meaning. Stochastic grammar reflects these features most effectively, by applying probability to each grammar [6,7,8].

Previous researches have been conducted to recognize all actions after the composition of some patterns to produce each grammar class. They need intricately huge labors to extract features of every object for its recognition and to artificially grammaticize distinctive ingredients. However, computers can construct standard patterns and automatically accumulate knowledge using an existing technique. A research has been actively performed on how an unknown pattern is recognized through accumulated knowledge. Syntactic recognition through grammatical inference is applicable to this case [9]. An application of grammatical inference was implemented in the field of music processing for modeling musical style. The models were used to generate and classify music by style [10].

Stochastic grammar inference is a recognition method that satisfies grammatical inference and stochastic grammar. This has already been implemented in several studies, such as normal or abnormal chromosome recognition [7], digit and shape recognition [11], text and speech recognition [12, 13] etc. Our study is an attempt to apply stochastic grammar inference to human action recognition.

2.2 Theoretical Concepts and Notations

In this subsection we introduce the inference of stochastic grammar method to seek the probability value of each production with given grammar and learning patterns. The given learning patterns are composed of subpatterns that respectively produce different grammar and probability value of each grammar.

Let's consider an M-class problem characterized by the stochastic grammars G_{sk}□$(N_k, \sum_k, P_k, D_k, S_k)$ for k□1, 2, ..., M. N_k is a finite set of nonterminals, $\sum_k$ is a finite set of terminals, P_k is a finite set of productions, D_k is a set of probability values of the production to be assumed, and S_k means the starting symbol [6]. In this paper, we define grammars for the actions of human upper body such as follows.

G_{sk}□(N, Σ, P, D_k, S), $k \in \{1...M\}$, N□{S, R, L, U, D}, Σ □{right, left, up, down},

P = { S -> right R	R -> right R	L -> right R	U -> right R	D -> right R
S -> left L	R -> left L	L -> left L	U -> left L	D -> left L
S -> up U	R -> up U	L -> up U	U -> up U	D -> up U
S -> down D	R -> down D	L -> down D	U -> down D	D -> down D
	R -> right	L -> right	U -> right	D -> right
	R -> left	L -> left	U -> left	D -> left
	R -> up	L -> up	U -> up	D -> up
	R -> down	L -> down	U -> down	D -> down }

To estimate D_k, the probability P_{kij} associated with the production $A_i \rightarrow \beta_j$ in G_{sk} must be obtained for each learning pattern set X of same actions. It is approximated by the relation ;

$$estimated \quad P_{kij} = \hat{} P_{kij} = \frac{n_{kij}}{\sum_r n_{kir}} \tag{1}$$

In equation (1) n_{kij} means the total average number of times when $A_i \rightarrow \beta_j$ in G_{sk} is used to all the learning patterns. It is obtained by equation (2).

$$n_{kij} = \sum_{x_h \, in \, X} n(x_h) \bullet p(G_{sk} / x_h) \bullet N_{kij}(x_h) \tag{2}$$

In equation (2) $n(x_h)$ is the frequency of all patterns occurred in X, $N_{kij}(x_h)$ is the number of times that $A_i \rightarrow \beta_j$ is used when a pattern of x_h is parsed. $p(G_{sk}/x_h)$ means the probability with which a pattern of x_h is produced from G_{sk}. $\sum_r n_{kir}$ in (1) is computed over all productions in G_{sk} that have the same A_i [6].

If D_k, for k=1…M, is obtained, the inferring step of the stochastic grammar as the learning step is completed, and the recognition step to recognize some arbitrary patterns can be performed.

3 Data Representation for Human Action Sequence

A representation for recognition of human action is developed. To satisfy the specification of stochastic regular grammar defined in this paper, raw data are converted to a suitable format using a preprocessing step.

3.1 Data and Data Aquisition

We utilized STABIL++ [14] to detect and track head and arm positions in video sequences. STABIL++ produces 3D positions of 11 color markers on each articulation, trunk, and head as in Fig.1. We call each color marker as a node. In our system, we use only 8 nodes such as 2 head nodes and 6 articulation nodes for 2 arms. Fig.2 shows an example of video data sequence: an action of left arm turning forward.

Data of node's movement is a sequence of the pair of x, y, z position, (x,y,z) on the world coordinates.

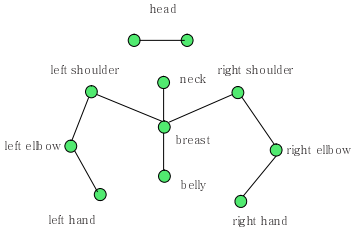


Fig. 1. Position of color markers.

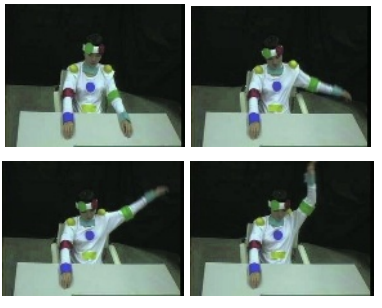


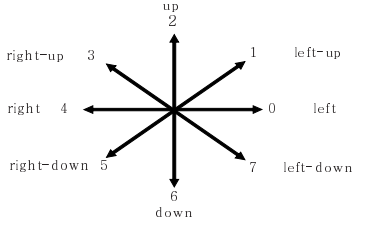
Fig. 2. Example of video data sequence.

3.2 Plane Projection of Quantized Data and 4-Connected Chain Coding

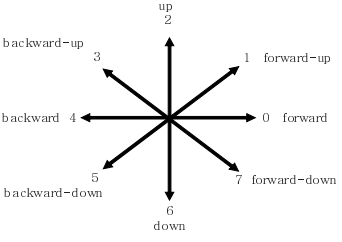
(x,y,z)s of articulation are used after their quantization at intervals of 3 cm so that very small motions in an action are disregarded. After converting quantized data into projection values in each xy, zy plane, we code them in 8-connected chain codes. The 8-connected chain coding of the projected data to xy plane has the meaning of left, left-up, up, right-up, right, right-down, down as in Fig.3(a). The 8-connected chain coding of the projected data to zy plane has the meaning of forward, forward-up, up, backward-up, backward, backward-down, down as in Fig.3(b). For example, the left window of Fig.3(c) shows the projection of a node movement to xy plane, and the right window is its projection to zy plane for an action in Fig.2.

After projection, a subpattern sequence in 8-connected chain codes are transformed into one in 4-connected chain codes. Too many productions are created by using 8-connected chain codes. It occurs high frequency of transition to next state and may decrease the recognition rate. To prevent it, we apply 4-connected chain coding. The 4-connected chain coding of the projected data to xy plane has the meaning of left, up, right, down. The 4-connected chain coding of the projected data to zy plane has the meaning of forward, up, backward, down. For example, forward-down in 8-

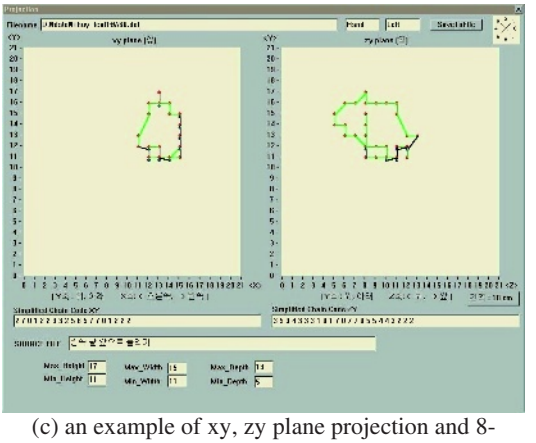
connected chain coding is transformed to forward and down in 4-connected chain code using the transformation rule [7]. The format of subpattern sequence will be indicated as (code code_count)s. “code” denotes each direction. “code count” refers to the repeating frequency of codes. An example of the subpattern sequence is shown in Table 1.



(a) xy plane : 8-connected chain code



(b) zy plane : 8-connected chain code



(c) an example of xy, zy plane projection and 8-connected chain coding

Fig. 3. 8-connected chain code and result after plane projection and 8-connected chain coding.

Table 1. An example of subpattern sequences for an action in Fig.2.

node = left hand plane = xy	left 12 up 11 right 4 up 5 right 9 down 13 left 2 down 1 left 5 down 2 left 3 up 13 left 1 up 1 #
node = left hand plane = zy	backward 23 up 16 forward 23 down 14 backward 6 down 2 backward 8 up 7 backward 2 up 1 backward 6 down 2 backward 15 up 3 backward 2 up 5 backward 2 up 9 forward 2 up 1 forward 2 up 1 forward 15 down 1 forward 2 down 2 forward 4 down 10 backward 24 up 2 backward 2 up 13 forward 3 up 2 forward 22 down 12 forward 2 down 5 #

3.3 Specification of Stochastic Regular Grammar

In our work, we represent 60 types of human actions using stochastic regular grammar. We estimate a stochastic value for each production of the regular grammar. Thus, M grammars that classify M human actions are expressed as follows. G_{sk_xy} is the grammar for the pattern projected to xy plane.

$$G_{sk\_xy} \square (N_{xy}, \Sigma_{xy}, P_{xy}, D_{k\_xy}, S), \quad k \in \{1...M\},$$

where $N_{xy} \square \{S, R, L, U, D\}$, $\Sigma_{xy} \square \{right, left, up, down\}$

G_{sk_zy} is the grammar for the pattern projected to zy plane.

$$G_{sk\_zy} \square (N_{zy}, \Sigma_{zy}, P_{zy}, D_{k\_zy}, S), \quad k \in \{1...M\},$$

where $N_{zy} \square \{S, B, F, U, D\}$, $\Sigma_{zy} \square \{backward, forward, up, down\}$

Fig.4 shows a state diagram of automata for the pattern projected to xy plane.

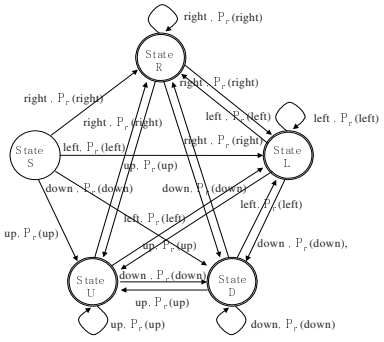


Fig. 4. State diagram of finite automata.

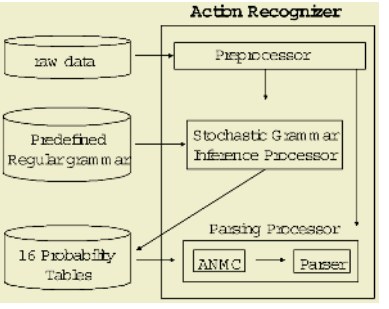


Fig. 5. The action recognizer.

4 Overview of Action Recognizer and Its Main Functions

Our action recognizer system is composed of a preprocessor, a stochastic grammar inference processor and a parsing processor. The preprocessor quantizes (x,y,z) data for input action and makes two sequences of 4-connected chain code for xy and zy projection planes. The stochastic grammar inference processor takes learning data and produces 16 probability tables by using the predefined regular grammar. Parsing processor takes subpattern sequences and searches a class with the highest stochastic value through the related stochastic information of the table. Fig.5 shows the structure of action recognizer. ANMC(Approximate Node Movement Classifier) is a preclassifier for searching action classes which have the same movement complexity. In this case movement complexity is defined as the combination of moving nodes.

4.1 Generation of Probability Tables

The inference step of stochastic regular grammar is applied to 8 nodes and for each projection plane. So, 16 probability tables are derived from the inference of the stochastic regular grammar.

The probability value for each production is obtained as shown in Table 2. It is an example of one node. The productions are replaced with the Confrontation Rules defined to follow the method of P_{kij}. The subscript "k" means the position of the grammar class, the subscript "i" means the subscript on the left hand side, and the subscript "j" means the subscript on the right hand side. Confrontation Rules are as follows.

A1 = S		
A2 = R	β1 = right R	β5 = right
A3 = L	β2 = left L	β6 = left
A4 = U	β3 = up U	β7 = up
A5 = D	β4 = down D	β8 = down

Some examples of the probability values practically estimated are shown in the column D_{44} of Table 2. D_{44} means the probability value of productions estimated after

Table 2. A probability table of left hand node for projection xy plane.

	Productions	Replaced Productions	D_1	D_2	……	D_{44}	……	D_m
1	S -> right R	A1 -> β1	P_{111}	P_{211}		0.12217		P_{m11}
2	S -> left L	A1 -> β2	P_{112}	P_{212}		0.56045		P_{m12}
3	S -> up U	A1 -> β3	P_{113}	P_{213}		0.30227		P_{m13}
4	S -> down D	A1 -> β4	P_{114}	P_{214}		0.01511		P_{m14}
5	R -> right R	A2 -> β1	P_{121}	P_{221}		0.69052		P_{m21}
…								
35	D -> up	A5 -> β7	P_{157}	P_{257}		0.00038		P_{m57}
36	D -> down	A5 -> β8	P_{158}	P_{258}		0.05257		P_{m58}

learning the action in Fig.2. This example shows the estimated probability value to the learning patterns of left hand node obtained through the projection on the xy plane.

4.2 Classification and Parsing for Recognition

When an action pattern is given, 16 subpattern sequences for 8 nodes on 2 projection planes are obtained after preprocessing. The parsing processor takes subpattern sequences as input, inspects moving nodes by ANMC and decides action classes that are compared. Then the nodes that are moved in action for each plane are calculated according to the "Multiplication Law of Probability", because all movements of nodes are statistically independent. The result probability for one action is calculated as follows. '$Node_i$' denotes a node probabilitiy.

$$P(Node_i \cap Node_j \cap \ldots \cap Node_k) = P(Node_i) \times P(Node_j) \times \ldots \times P(Node_k),$$
$$\text{for any } i,j,k \in \{1 \ldots 16\} \text{ and } i \neq j \neq k$$

The parser estimates every probability for each action class that is comparable, and looks for an action class with the highest probability among them.

5 Experimentation and Results

In order to confirm the effectiveness of our method, three types of experiments are carried out. We show the experiment data, kind of experiments and recognition performance of each experiment in our system.

5.1 Experimental Data

900 human upper body actions of 60 types that 3 persons gestured were recorded in STABIL++ system. 490 action data are used for learning and 410 action data are used for testing. Practically, the data for recognition to analyze actions are composed of the movements of head, bodies, arms, and of other compound body movements as in Table 3. These kinds of actions are selected with the reference to the data for human action analysis studies [15]. For example, head movements are hang down head and to the former place, raise head upward, turn head to the right and to the former place.

Table 3. Summary of human upper body actions.

Complexity	Body Part	Movement type	Direction	No. of Actions
Primitive	Head	hang down, raise up, turn, bend, rotate	up, down, right, left, forward, backward	8
	Trunk	bend, turn, lean, shake	right, left, forward, backward	9
	Right hand or arm	contacting, turn, raise	up, down, right, left, forward, backward	14
	Left hand or arm	contacting, turn, raise etc	"	14
Combination	Both hands or arms	raise, turn, fold, cross	"	8
	Others		"	7

5.2 Experimental Result

Table 4 shows the results that have been obtained from 3 kinds of experiments with our action recognizer. Experiment 1 of 8 action types of head shows the recognition rate of 93.8%. In this experiment we don't use the ANMC because head actions have all same movement complexity. 40 action data are used for learning and 32 action data for testing. Experiment 2 of 60 action types of head and body without ANMC shows the recognition rate of 64.6%. Experiment 3 of 60 action types of head and body with ANMC shows the recognition rate of 84.9%. 490 action data are used for learning and 410 action data for testing in experiment 2 and 3. The recognition rate is estimated as Recognition Rate=Number of Correctly Recognized Actions/Total Number of Actions.

Table 4. Experimental results.

Experiment	Experimental Contents	Learning Data	Recognition Rate
1	8 action types of head without ANMC	40	93.8 % (30/32)
2	60 action types of head and body without ANMC	490	64.6 % (265/410)
3	60 action types of head and body with ANMC	490	84.9 % (348/410)

Our recognition method achieved comparatively high recognition rate for only one node, where the movement complexity is 1 such as in Experiment 1. Actions in Experiment 2 and 3 have the movement complexity from 2 to 8. Nevertheless, we achieved comparatively high recognition rate in Experiment 3. A pre-classification technique of ANMC causes the recognition rate to be improved.

6 Conclusions

In this paper we proposed a recognition model to understand human upper body actions using stochastic grammatical inference method. Grammatical inferring has been left unexplored and referred to as the further study [3].

3D data sequence of human actions are encoded into 4-connected chain codes, and projected to xy and zy plane. These sequences are processed as the input of the stochastic recognizer. They are used in the learning step for building the probability

tables. Using these tables in the recognition step, each action is classified into the befitting class of human action.

Our scheme is suitable not only for simple actions composed of single articulation movement, but also for complex actions composed of several articulation movements. We have showed the possibility of autonomous learning from predefined human action patterns. In our experiments, 93.8% recognition rate of 8 action types of human head and 84.9% recognition rate of 60 action types of human upper body were achieved.

The contact of human hands to another body part has much importance for the analysis of nonverbal actions of human [5,15]. However, the recognition system through the general inference method by the stochastic grammar doesn't yet reflect the characteristics of nonverbal actions. The task for the future study may be the work to extend the inference method in this paper to reflect peculiarities of nonverbal actions.

References

1. Becker, D.A. 'Sensei: A Real-Time Recognition, Feedback and Training System for T'ai Chi Gestures,' MIT Media Lab Perceptual Computing Group TR 426 (1997)
2. Moeslund, T.B and Granum, E., 'A Survery of Computer Vision-Based Human Motion Capture,' Computer Vision and Image Understanding, vol. 81, No.3 (2001) 231-268 (2001)
3. Inanov, Y.A.: Application of stochastic grammars to understanding action, MIT thesis (1998)
4. Hong, P., Turk, M., and Huang, T., 'Gesture Modeling and Recognition Using Finite State Machines,' Int'l Conference on Gesture Recognition, Grenoble, France (2000)
5. Frey, S.: 'Das Berner System zur Untersuchung nonverbaler Interaktion,' Methoden der Analyse von Face-to-Face-Situationen, Stuttgart, Metzler, (1981) 203-236
6. Gonzalez, R.C. and Thomason, M.G.: Syntactic Pattern Recognition, Addison-Wesley Publishing Company (1978) 177-270
7. Fu, K.S.: Syntactic Methods in Pattern Recognition, Academic Press (1974) 54-55, 124-229
8. Doest, H.: 'Stochastic Grammars: Consistency and Inference', Beoorderlingscommissie, (1994), 53-84
9. Gronfors, T. and Juhola, M.: 'Experiments and comparison of inference methods of regular grammars', IEEE Trans. on Systems, Man and Cybernetics, Vol.22(4), (1992) 821-826
10. Cruz-Alcazar, P.P. and Vidal-Ruiz, E.: 'Modeling musical style using grammatical inference techniques: a tool for classifying and generating melodies', Third International Conference on Web Delivering of Music, (2003) 77-84
11. Vidal, E., 'Application of the error-correcting grammatical inference algorithm (ECGI) to planar shape recognition', IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, (1993) 24/1 - 24/10
12. Atwell, E., et al, 'Multi-level disambiguation grammar inferred from English corpus, treebank, and dictionary', IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, (1993) 9/1 - 9/7
13. Galiano, I. and Segarra, E., 'The application of k-testable languages in the strict sense to phone recognition in automatic speech recognition', IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, (1993) 22/1 - 22/7
14. Munkelt, O., et al., 'A model driven 3D image interpretation system applied to person detection in video images,' 14th ICPR 98 (1998) 70-73
15. Bull, P.E.: Posture and Gesture, Pergamon Press (1990) 163-187