Indexicality and Dynamic Attention Control in Qualitative Recognition of Assembly Actions

Yasuo Kuniyoshi[1] *and Hirochika Inoue*[2]

[1] Autonomous Systems Section, Intelligent Systems Division, Electrotechnical Laboratory, 1-1-4 Umezono, Tsukuba-shi, Ibaraki 305, JAPAN
[2] Department of Mechano-Informatics, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, JAPAN

Abstract. Visual recognition of physical actions requires temporal segmentation and identification of action types. Action concepts are analyzed into attention, context, and change. Temporal segmentation is defined as a context switch detected by a switching of attention. Actions are identified by detecting "indexical" features which can be quickly calculated from visual features and directly point to action concepts. Validity of the indexicality depends on the attention and the context. These are maintained by three types of attention control: spatial, temporal and hierarchical. They are combined by a mechanism called "attention stack", which extends at important points and winds up elsewhere. An action recognizer built upon the framework successfully recognized human assembly action sequences in real time and output qualitative descriptions of the tasks.

1 Introduction

Intelligent systems in multi-agent environment must react to other agents' actions. In such cases, visual recognition of actions is indispensable. The recognition process must operate in real-time, generating semantic information suitable for reasoning processes or matching against stored knowledge.

Research on motion understanding has offered various methods to estimate 3D motion parameters of various objects including human bodies [1, 2]. But it is not clear how to extract semantics from these parameters. Moreover, bodily motion parameters are not sufficient nor of primary importance for action recognition. Early attempts to extract sematic information from motion pictures dealt with simple animations [3, 4]. Real-time recognition of general actions in the real world is still an open problem.

Recently, many ill-posed vision problems have been reformulated and solved by assuming an active observer [5] and considering behavior context [6, 7]. This approach should fit well to qualitative action recognition.

In this paper, we propose and test a new method for real-time visual recognition of simple assembly tasks performed by human workers. The core of the method is two fold: (1) Define a set of simple "indexical" features which can be quickly calculated from visual features. They directly point to action concepts under certain conditions. (2) Employ spatial/temporal/hierarchical attention control and context memory to maintain the indexicality of extracted features.

2 Qualitative Recognition of Human Action Sequences

The objective of qualitative action recognition is to generate symbolic descriptions of every action and every related change in the environmental state by watching continuous performance of purposeful human actions.

The example problem treated in this paper is stated as follows:

Input: A human worker constructs various structures with blocks. One structure is assembled in one task. Each task is carried out only once from start to finish without a pause. The system observes it by stereo vision.

Output: Symbolic assembly plans. Each plan is a sequence of assembly operators with intermediate state descriptions. It should be useful for reproducing the task by a robot [8].

Assumptions: Explicit cues for temporal segmentations and a priori knowledge about a specific task are not given. Classification of possible assembly operations is given. The worker uses only one hand and grasps only one block at a time. Spatial segmentation is tractable. Complete occlusion of blocks should not occur. "Messed up" situations such as collapses are not allowed.

The system must automatically determine the start and the end time point of every action (temporal segmentation), identify each temporal interval as either of known action class (action identification), extract and remember qualitative state descriptions at the segmentation points (state recognition). All these must be processed in real time. This constraint gives rise to the use of contextual information for active attention control while freeing the system from storing whole image sequences.

3 Indexicality in Classifying Actions

In order to define indexical features for assembly actions, we first analyze the structure of action concepts and then construct mappings from observable features to action concepts.

Conceptual Organization of Assembly Actions. Let W be a set of all possible episodes in the time-space of the physical world. Qualitative recognition of an action requires cutting a spatio-temporal region from W and identify as one of known action concept.

According to Hobbs [9], a set of concepts C is modelled as a quotient set of W by a certain equivalence relation $\sim_c$: $C = W/\sim_c$). A set of observable qualitative features F is also defined as a quotient set of W by an equivalence relation $\sim_o$: $F = W/\sim_o$.

We state that F has indexicality to C if and only if the observation process $\sim_o$ articulates the world W in the very same way as the concept formation process $\sim_c$ does.

Let us divide the concept formation process $\sim_c$ into three steps (note that all three are equivalence relations):

Attention: Specify "focused entities" (objects, geometrical features or locations) which serve as "supports" of motion/relation descriptions. Their identity must be maintained within each action interval.

Context: Invariant portion of motion: Relative position/velocity and contact state of focused entities.

Change: Qualitative change of relations (held, affixed) or relative position/orientation (on, coplanar) among focused entities. Equivalence relation among possible temporal changes in W. Meaningful only under specified context.

Physical actions in assembly tasks are roughly grouped by the "context" into "assembly motions": (1)"Transfer", with large movement of the hand in free space, and (2)"LocalMotion" with the hand movement near specific target objects. "LocalMotion" is further divided into three types of assembly motions: (1)"Approach", in which the fingers and the held object moves toward the target objects, (2) "Depart", the other way round, and (3) "FineMotion", in which the held object moves in contact with the target objects.

Indexical Features of Assembly Actions. Provided that the "attention" and the "context" are maintained, assembly actions belonging to each assembly motion are discriminated by the "change". Observable features which directly point to the "changes" can be defined in terms of temporal change of simple visual features.

Action concept of LocalMotion is stated as follows:

Attention: The hand, the target object and the held object, if any.
Context: The hand is moving slowly near (toward/away) the target object.
Change: Δholding(Hand, X) classifies PICK/PLACE/NO.

Let us define a 3D convex region called "target region" which covers the block to be picked up or the space in which the held one will be placed. Then, the indexical feature is defined as a change of the intersecting volume of the target region and any object. This feature can be detected by the following procedure: (1) Predict the location and size of the target region. (2) Project the 3D model of the target region onto each field of view of binocular stereo to define 2D attention regions. (3) For each view, take two gray level (possibly, color) snapshots of the attention region at temporal segmentation points before and after the current action. (4) Differentiate the snapshots and threshold the change of area into three cases, decreased/increased/no-change, which point to PICK/PLACE/NO[3]. Then verify the 3D location of the change by triangulating the center of mass of the changed areas.

Conceptual organization of an ALIGN operation (see Fig. 1) is as follows:

Attention: F1 of B1 (moved block), F2 of B2 (target block).
Context: (1) B1 and B2 are identified. (2) Bottom face of B1 and top face of B2 are "against". (3) B1 is moving in direction **m** so that F1 approaches the common plane with F2.
Change: $\neg$coplanar(F1, F2) $\rightarrow$ coplanar(F1, F2)

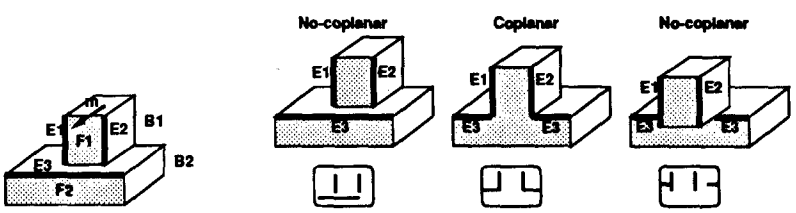

Fig. 1. ALIGN action (leftmost) and indexical features (in round boxes) of coplanar/no-coplanar relations.

In the context specified above, simple 2D visual features, namely, contour junction types shown in Fig. 1, directly point to coplanar/no-coplanar states. Tracking only the edges E1, E2, E3 and check for the junctions, coplanar states are immediately detected. The "change" feature can be quickly calculated by differentiating the coplanar/no-coplanar state at start/end points of the current action.

[3] To simplify the implementation, blocks are painted white and the background is black, in our experiment.

4 Dynamic Attention Control for Action Recognition

The indexicality of observable features is based on correctness of the "attention" and the "context".

Spatial attention control maintains the "attention" part. It detects the moving hand by temporal differentiation of images (pre-attentive), tracks the hand and the moving objects/edges to solve correspondence problem, and predicts a target region/edge by visual search procedure. The visual search starts by setting a stereo-pair of 2D regions on the "base" of attention maintained by tracking. The regions are progressively moved in the direction in which the base is moving until they hit an object or the worktable. By looking up the estimated 3D position into the environment model, a correct target region is determined. Contextual information is important here, eg. the environment model is updated by recognizing previous actions, and whether to set the target region on (in PICK) or above (in PLACE) the found object depends on which action is expected now.

Temporal attention control maintains the "context" part. When a shift of attention is detected by a visual search or a pre-attentive routine, current context is no longer valid, which means that the indexcality of currently selected features breaks. This context switch directly signals a temporal segmentation of an action. Different visual features are selected for monitoring, and tracking or visual search is invoked/stopped depending on the context. This is done by selecting nodes of the action model.

Hierarchical attention control stabilizes the overall recognition process. It extracts different types of visual features at different timings in parallel from superimposed regions. The regions and the features are organized in a hierarchy which corresponds to the levels of assembly motions. Very coarse features such as temporal differentiation of the whole view are always monitored, while fine features such as edge junctions are checked only in FineMotions. The hierarchical parallelism contributes to the robustness: Even when the hand suddenly moves off while a FineMotion, the system readily catches up with the gross motion. This control scheme is implemented as "attention stack" operations.

5 Experiments and Results

An experimental system [10] was built upon the presented theory. It succeeded in recognizing human assembly actions in real-time. Figure 2 shows a monitor display during the recognition of an arch construction task. The worker's hand is tracked by multiple visual windows and an indexical feature is extracted from the target region. Result of recognition is displayed at the bottom lines. The elapsed time for the whole task was 2 min. Figure 3 shows the coplanar detector at work. In this experiment, the detector was run continuously to repeatedly check for coplanar relations. The held block was moved continuously and the detector reported coplanar/no-coplanar discrimination at a speed of 1 Hz. Other tasks recognized successfully by the system include; (1) pick and place with ALIGN operation, (2) tower building, (3) inverted arch balanced on a center pillar, (4) table with four legs.

Acknowledgement

The authors wish to express their thanks to Dr. Masayuki Inaba for his suggestions and contribution to the base system, and Mr. Tomohiro Shibata for implementing the coplanar detector.

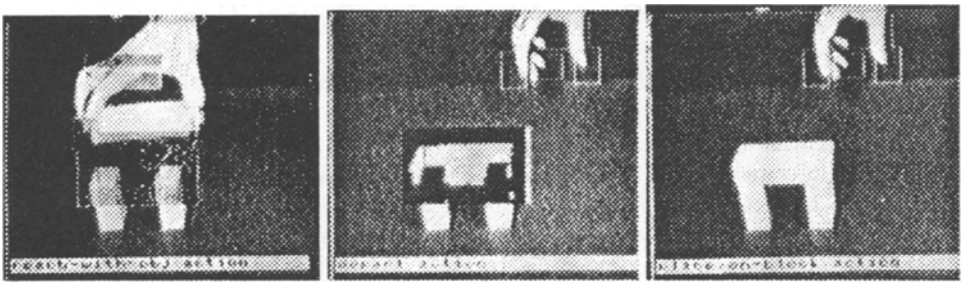

Fig. 2. Recognizing the task "Build Arch": (1) Target region defined. (2) Differentiation. (3) Identified action type "place-on-block" displayed.

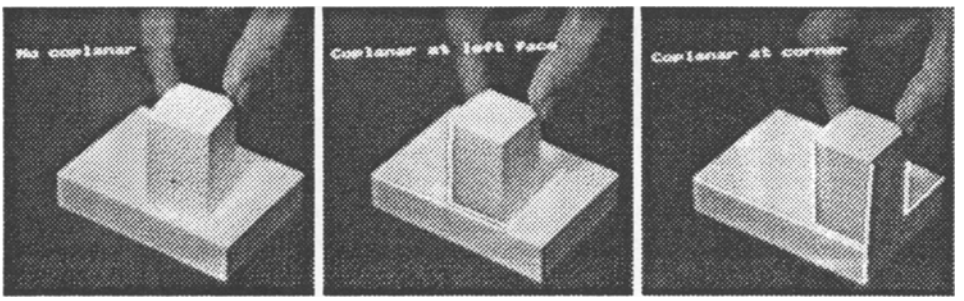

Fig. 3. Recognizing an ALIGN action. Indexical features displayed as white lines.

References

1. J. O'rourke and N. J. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Trans.*, PAMI-2(6):522–536, 1980.
2. M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *Proc. of IEEE Conf. CVPR*, 664–665, 1991.
3. S. Tsuji, A. Morizono, and S. Kuroda. Understanding a simple cartoon film by a computer vision system. In *Proc. IJCAI5*, 609–610, 1977.
4. R. Thibadeau. Artificial perception of actions. *Cognitive Science*, 10(2):117–149, 1986.
5. J. Aloimonos and I. Weiss. Active vision. *Int. J. of Computer Vision*, 333–356, 1988.
6. D. H. Ballard. Reference frames for animate vision. In *Proc. IJCAI*, 1635–1641, 1989.
7. S. D. Whitehead and D. H. Ballard. Learning to perceive and act. Technical Report 331, Computer Science Dept., Univ. of Rochester, Rochester, NY 14627, USA, June 1990.
8. Y. Kuniyoshi, M. Inaba, and H. Inoue. Teaching by showing: Generating robot programs by visual observation of human performance. In *Proc. ISIR20*, 119–126, 1989.
9. J. R. Hobbs. Granularity. In *Proc. IJCAI*, 432–435, 1985.
10. Y. Kuniyoshi, M. Inaba, and H. Inoue. Seeing, understanding and doing human task. In *Proc. IEEE Int. Conf. Robotics and Automation*, 1992.

This article was processed using the LaTeX macro package with ECCV92 style