EXTENDED ABSTRACT

The theory of Therbligs: A compositional approach to incremental robot learning

Yiannis Aloimonos
Computer Vision Lab
Perception for Robotics group
prg.cs.umd.edu
University of Maryland

Therbligs comprise a system for analyzing the motions involved in performing a task. Frank and Lillian Gilbreth invented and refined this system, roughly between 1908 and 1924. Their motivation was to reduce unnecessary motions for the worker and increase productivity. They arrived at this general system through systematic examination of workers in factories, in a variety of settings. The name Therblig is Gilbreth spelled backwards.

The Gilbreths observed empirically that no matter what the activity was, there always existed a number of **primitives**, or basic movements that made up any (manipulation) activity. For example, during an activity the hand is moving (while empty) towards an object. This whole motion is one Therblig. Then the hand grasps an object, and this is another Therblig. Subsequently, the hand is moving (while loaded) towards some location, making another Therblig. The Gilbreths came up with 15 Therbligs which later were extended to 18 by other authors. They include: **TE** -Transport empty (hand moving with nothing in it), **TL** - Transport loaded (hand moving with something in it), **G** – grasp an object/tool, **H** – Hold, **R** – Release, **PP** – Preposition, **P** – Position, **U** – Use**, A** – assemble, **DA** – disassemble, and few more having to do with searching, finding and delays. If we consider the objects involved, then the therbligs take them as arguments, e.g. G(noun) means Grasp object denoted by “noun”.

For example, if someone is making a bowl of cereals and bread with milk for breakfast, the whole activity can be turned into a string of Therbligs: TE (Plate) - G (Plate) - TL (Plate, Table) - R (Plate) - TE (Fridge) - U (Fridge) - TE (Container) - G (Container) - TL (Container, Table) - R (Container) - TE (Fridge) - U (Fridge) - TE (Milk) - G (Milk) - TL (Milk, Table) - R (Milk) - U (Cabinet) - TE (Box) - G (Box) - TL (Box, Table) - U (Cabinet) - TL (Cereal, Table) - R (Cereal) - TE (Cup) - G (Cup) - TL (Cup, Table) - R (Cup) - P (Cereal) - H (Cereal) - TE (Bread) - G (Bread) - TL (Bread, Bowl) - R (Bread) - TE (Box) - G (Box) - H (Box) - TL (Cereal, Box) - U (Box) - U (Cabinet) - TL (Box, Cabinet) - R (Box) - U (Cabinet) - TE (Milk)-….Given this string in Therblig language, it is easy for a semantic parser to produce the sequence of actions taking place ((i.e. take, put, open, close, pour, etc), assuming we can obtain more information about the Therblig **U** (context dependent – **U** could be push, pull, press, move, cut, slice..).

For a robot to learn something, it means it learns something new to do, i.e. some action. But any possible action is a sequence of Therbligs. Thus the Therbligs become a universal language for action and learning becomes a well defined problem. We need compilers that can translate a visual action (the video of an action) into Therblig language (learning from vision) and compilers that translate action given in Natural Language into Therblig language (learning from language). I will describe our efforts to address these questions using Transformer Networks and show recent experimental results.