Imitation Learning of Team-play in Multiagent System based on Hidden Markov Modeling

Itsuki Noda
Cyber Assist Research Center, AIST and PRESTO, JST, Japan
i.noda@aist.go.jp

Abstract

This paper addresses agents' intentions as building blocks of imitation learning that abstract local situations of the agent, and proposes a hierarchical hidden Markov model (HMM) to represent cooperative behaviors of teamworks. The key of the proposed model is introduction of gate probabilities that restrict transition among agents' intentions according to others' intentions. Using these probabilities, the framework can control transitions flexibly among basic behaviors in a cooperative behavior.

1 Introduction

Imitation learning is considered to be a method to acquire complex human and agent behaviors and as a way to provide seeds for further learning [KI93; Sch99; MK98]. While those studies have focused on imitating behaviors of single agents, few works address imitation for teamwork among multiple agents because the complexity of the world state increases drastically in multi-agent systems. On the other hand, stochastic models like hidden Markov models (HMM) have been studied as tools to model and to represent multi-agent/human interactions [ORP00; IB99]. It is, however, hard to apply these stochastic models to imitate teamworks by observation because of the complexity of the model of multiple agents. This study focuses upon *intentions* of agents as building blocks of an abstract state of the local world for the agent in order to overcome the problem. Using *intention*, I formalize teamwork and propose a hierarchical hidden Markov model for imitation learning of teamwork.

2 Teamwork and Imitation

2.1 Intention and Play

We suppose that an *intention* is a short-term idea to achieve a certain condition from another condition. For example, in soccer, the intention 'to guide a ball in a certain direction' is an idea to move to a certain direction with the ball. We assume that an *intention* is an individual idea; therefore, an agent does not pay attention to others' efforts to achieve their intention.

A *play* is postulated as a sequence of atomic actions to achieve a single *intention*. The *play* is a basic building block of overall behavior of agents. For example, in soccer, a 'dribble' is a *play* to achieve the *intention* 'guide a ball in a certain direction', which consists of atomic actions like 'turn', 'dash', 'kick', and so on. A play for the intention is also an individual behavior without collaboration with other agents because an intention is an individual idea. As shown below, an intention and the corresponding play are used as a main trigger to synchronize team-plays among multiple agents. This means that the intention is treated as a kind of partial condition of the world.

2.2 Team-play

We suppose that *team-play* is a collection of plays performed by multiple agents to achieve a certain purpose. As mentioned in the previous section, an intention is an individual idea. This means that multiple agents who do not change their intentions can not perform a *team-play* because they have no way to synchronize their plays. Instead, we assume that they can synchronize their plays by changing their intentions according to situations of environments and intentions of other agents. For example, in soccer, when two players (passer and receiver) guide a ball by dribble and pass, players will change their intentions as follows:

passer	**receiver**
dribbling	:
↓	supporting
seek-receiver	↓
↓	free-run
pass	↓
:	receive

In this example, the passer and the receiver initially have intentions 'dribbling' and 'supporting', respectively. Then, the passer changes the intention to 'seek-receiver', followed by the receiver's change to 'free-run', the passer's change to 'pass', and so on. Play synchronization is represented as conditions when agents can change the intention. In the example, the passer changes its intention from 'seek-receiver' to 'pass' when the teammate's intention is 'free-run'.

2.3 Imitation Learning of Team-play

The imitation learning of a teamplay is formalized as follows: (1) **Observation:** to observe behaviors of mentor agents and estimate what intention each agent has at each time step. (2) **Extraction:** to extract conditions prevailing when each agent

changes intentions. A condition is represented as a conjunction of others' intentions. (3) **Generation:** to generate a sequence of intentions according to changes of environment and others' intentions. In the second step of this process, the *intention* plays an important role: that is, conditions of changes of intentions. As described in Section 2.1, we consider that *intention* can represent world conditions. In addition to it, we use only *intention*s to construct rules for agents to change their *intention*s.

3 Hierarchical Hidden Markov Model for Agents

3.1 Single Behavior Model

We formalize behaviors of a basic *play* m performed by a single agent as a Moore-type HMM as follows:

$$\mathbf{HMM^s}_m = \langle \boldsymbol{S}_m, \boldsymbol{V}_m, \boldsymbol{P}_m, \boldsymbol{Q}_m, \boldsymbol{R}_m \rangle ,$$

where $\boldsymbol{S}_m = \{s_{mi} | i \in \boldsymbol{S}_m\}$ is a set of states, $\boldsymbol{V}_m = \{v_k\}$ is a set of a pair of sensor value and action commands which are used as outputs from the state, $\boldsymbol{P}_m = \{p_{mij} | i, j \in \boldsymbol{S}_m\}$, $\boldsymbol{Q}_m = \{q_{mi}(v) | i \in \boldsymbol{S}_m, v \in \boldsymbol{V}_m\}$, and $\boldsymbol{R}_m = \{r_{mi} | i \in \boldsymbol{S}_m\}$ are probability matrixes of state transition, state-output, and initial state, respectively.

3.2 Cooperative Behavior Model

As discussed in the previous section, we consider that team-play consists of a sequence of intentions of multiple agents. This means that cooperative behavior of a single agent in a team of agents is considered as transitions among several basic plays (**HMM**$^{\mathbf{s}}$). Therefore, we formalize cooperative behavior as the following modified Mealy-type HMM,

$$\mathbf{HMM^c} = \langle \boldsymbol{M}, \boldsymbol{U}, \boldsymbol{E}, \boldsymbol{F}, \boldsymbol{G}, \boldsymbol{H} \rangle ,$$

where $\boldsymbol{M} = \{\mathbf{HMM^s}_m\}$ is a set of basic plays and $\boldsymbol{U} = \{u_k\}$ is a set of output from the model (normally, same as $\boldsymbol{M}$); $\boldsymbol{E} = \{e_m | m \in \boldsymbol{M}\}$ is a set of initial play probabilities, $\boldsymbol{F} = \{f_{min} | m \in \boldsymbol{M}, i \in \boldsymbol{S}_m, n \in \boldsymbol{M}\}$ is a set of exiting probabilities from plays, and $\boldsymbol{G} = \{g_{mnj} | m \in \boldsymbol{M}, n \in \boldsymbol{M}, j \in \boldsymbol{S}_n\}$ is a set of entering probabilities to plays. Also, $\boldsymbol{H} = \{h_{mn}(u) | m \in \boldsymbol{M}, n \in \boldsymbol{M}, u \in \boldsymbol{U}\}$ is a set of gate probabilities between plays. Formally, these probabilities are defined as: $e_m = Pr(R_m^{\langle 0 \rangle})$, $f_{min} = Pr(R_n^{\langle t \rangle} | s_{mi}^{\langle t \rangle})$, $g_{mnj} = Pr(s_{nj}^{\langle t \rangle} | R_m^{\langle t-1 \rangle})$, and $h_{mn}(u) = Pr(u^{\langle t-1 \rangle} | R_m^{\langle t-1 \rangle}, R_n^{\langle t \rangle})$.

Using these probabilities, an actual probability from state i in play m to state j in play n is calculated as

$$Pr(s_{nj}^{\langle t \rangle} | s_{mi}^{\langle t-1 \rangle}) = \begin{cases} f_{mim} p_{mij} \; ; m = n \\ f_{min} g_{mnj} \; ; m \neq n \end{cases} .$$

3.3 Joint-Behavior Model

Finally, we coupled multiple **HMM**$^{\mathbf{c}}$s, each of which represents the behavior of an agent. Coupling is represented by gate probabilities $\boldsymbol{H}$. For example, when agent X and agent Y are collaborating with each other, the gate probability $h_{mn}(u)$ in **HMM**$^{\mathbf{c}}$ for agent X indicates the probability that agent Y is performing play u at time t when agent X changes the play from m to n during time $t \rightarrow t+1$. Using the gate probability, the agent calculate a likelihood of the state s_{nj} at a certain time $t+1$ according to the following equation:

$$L(s_{nj}^{\langle t+1 \rangle}) = \begin{cases} f_{mim} p_{mij} q_{mj}(\hat{v}^{\langle t+1 \rangle}) & m = n \\ f_{min} g_{mnj} h_{mn}(u^{\langle t \rangle}) q_{nj}(\hat{v}^{\langle t+1 \rangle}) & m \neq n \end{cases} \quad (1)$$

, where $\hat{v}^{\langle t+1 \rangle}$ is a partially observed output value in v at time $t+1$.

4 Experiments

I applied the framework to collaborative play of soccer. The demonstration by mentors is dribble and pass play as shown in Fig. 1: A player starts to dribble from the center of the left half field and brings the ball to the right half. At the same time, another player runs parallel along the upper (or lower) side of the field supporting the dribbling player. Then, the first player slows to look-up the second player; it then passes the ball to that player. Simultaneously, the second player starts to dash to the ball and dribbles after receiving the ball. After the pass, the first player exchanges roles with the teammate so that it becomes a supporting player for the second player.

To imitate this demonstration, I trained six **HMM**$^{\mathbf{s}}$s to model 'dribble', 'slow-down and look-up', 'pass', 'free-run', 'chase-ball', and 'support'. Each of **HMM**$^{\mathbf{s}}$s has 5 states. The output of these **HMM**$^{\mathbf{s}}$s consists of local situations (the relative position and the velocity to the ball) and agent's actions ('turn', 'dash', 'small-kick', 'long-kick', 'trap', and 'look'). Note that there is no information about others' situations for output of **HMM**$^{\mathbf{s}}$s. As described in Section 2.2, others' situations are taken into account during the **Extraction** phase in learning.

Two **HMM**$^{\mathbf{c}}$s for agent X (the first player) and Y (the second player) are constructed after the training of the **HMM**$^{\mathbf{s}}$s. Then, the learner observes behaviors of the mentor and adjusts probabilities of the **HMM**$^{\mathbf{c}}$s.

Figure 2 shows result of observation and estimation. This figure shows the relative likelihood of each play state for each agent at each timestep estimated by **Observation** phase. In this figure, there are 12 rows of small squares: upper 6 rows correspond 6 plays of the first player (agent X), and the rest 6 are plays for the second player (agent Y). Each row corresponds to a play D, K, P, F, C, and S, each of which means 'dribble (D)' , 'slow-down and look-up (K)', 'pass (P)', 'free-run (F)', 'chase-ball (C)', and 'support (S)', respectively. In each row, a column consists of 5 small squares each of which corresponds a state of **HMM**$^{\mathbf{s}}$for one of the 6 plays at a certain timestep. The ratio of black area in the square indicates the relative likelihood with which the state of the **HMM**$^{\mathbf{s}}$is active at the timestep. Columns are aligned along with time. So, a horizontal line of squares means changes of likelihood of a state of **HMM**$^{\mathbf{s}}$. From this figure, we can see that the learner estimates that the agent X starts the play with the intention of 'dribble', followed by 'slow-down', 'pass' and 'support', while the player Y starts 'support' play, followd by 'chase-ball' and 'dribble' plays.

After the training by the **Observation**, the learner can generate behaviors similar to the demonstration by using the acquired probabilities of the **HMM**$^{\mathbf{c}}$as shown in Fig. 3. This

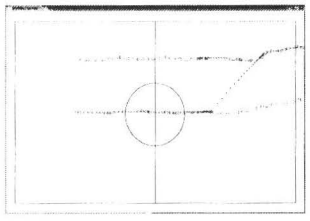

Figure 1: Exp. 3: Dribble and Pass Play by Mentor

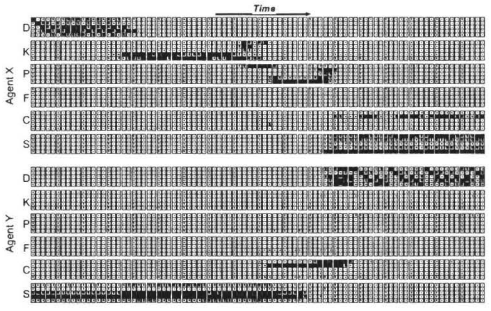

Figure 2: Exp. 3: Result of Recognition of Mentor's Behaviors

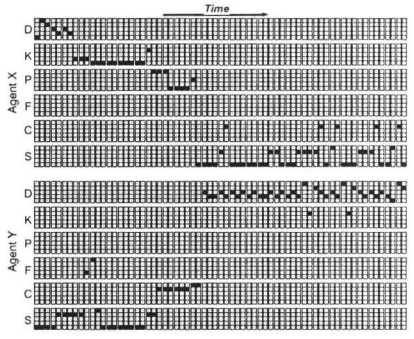

Figure 3: Exp.3: State Transitions Generated by Learned HMM

figure is constructed in the same way as Fig. 2, but only one square is filled in a timestep because the learner decides one of the possible states according to the likelihood shown in Eq. 1. In this example, although the learner sometimes generates wrong state transitions, for example a transition to states to the 'free-run' play in agent Y during agent X is doing 'slow-down', it recovers to the suitable transitions and continues to imitate the demonstrator. This shows robustness of the model against accidents. Because the model is coupled loosely with world and other's states by output probabilities of HMM, it can permit variation and misunderstanding of world and others' states.

5 Discussion

There are several works on coupling HMMs that can represent combinational probabilistic phenomena like multi-agent collaboration [JGS97; GJ97; JGJS99]. In these works, probabilistic relation among several HMMs (agents) are represented as state-transition probabilities, such that the amount of memory complexity increases exponentially. This is a serious problem for imitation learning because we assume that the number of examples for imitation is small. In our model, the relation among agents is represented by gate probabilities $\boldsymbol{H}$, in which others' states are treated as outputs instead of as conditions of state transition. Using them, the likelihoods of state-transitions are simplified as products of several probabilities (Eq. 1). In addition, detailed states of other agents are abstracted by play (intention). As a result, the number of parameters is reduced drastically, so that learning requires very small number of examples as shown in above examples. Although such simplification may decrease flexibility of representation as a probabilistic model, experiments show that the proposed model has enough power to represent team-play among agents.

Intention in the model brings another aspect to communication among agents. We assume that there are no mutual communication in the proposed model. However, we can introduce communication as a bypass of observation and estimation of other's intention (play). The proposed model will be able to provide criteria for when an agent should inform their intention to others by comparing agents' actual intentions and estimated intention of the agent itself by simulating its own $\mathbf{HMM}^{\mathbf{c}}$.

One important issue is the design of the intention. In the proposed model, intentions play various important roles like chanking of the actions and conditions of world state. Therefore, we must design intentions carefully so that team-plays can be represented flexibly.

References

[GJ97] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Machine Learning*, 29:245–275, 1997.

[IB99] Y. A. Ivanov and A. F. Bobick. Recognition of multi-agent interaction in video surveillance. In *International Conference on Computer Vision (Volume 1)*, pages 169–176. IEEE, Sep. 1999.

[JGJS99] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[JGS97] Michael I. Jordan, Zoubin Ghahramani, and Lawrence K. Saul. Hidden markov decision trees. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 501. The MIT Press, 1997.

[KI93] Y. Kuniyoshi and H. Inoue. Qualitative recognition of ongoing human action sequences. In *Proc. IJCAI93*, pages 1600–1609, 1993.

[MK98] Hiroyuki Miyamoto and Mitsuo Kawato. A tennis serve and upswing learning robot based on bi-directional theory. *Neural Networks*, 11:1331–1344, 1998.

[ORP00] Nuria M. Oliver, Barbara Rosario, and Alex Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.

[Sch99] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, Jun. 1999.