人　工　知　能　99−12
（１９９５．３．７）

強化学習を用いたサッカーゲームモデル

○久保正男（北大院）　　嘉数侑昇（北大）

Collective behaviorおよび強化学習に含まれる学習システムの問題の一つである，如何に現実的で有効な行動集合を設定するかという課題は，学習環境が複雑になればなる程困難を極める．つまり低レベルな行動集合で効果的に学習を行う為には十分な状態数が必要であり，しばしば学習時間が膨大な量になる．そこでオンラインでの強化学習を行いながら，かつ行動集合を更新する機構の構築を目的とする．また複雑な学習環境としてマルチエージェントシステムによるサッカーにアプローチする．

Soccer Game Model by Multi-Agent System with Reinforcement Learning

Masao KUBO*　　Yukinori KAKAZU**
*Division of Information Engineering
Hokkaido University
**Department of Precision Engineering,
Hokkaido University

This study theoretically concerns the exploitation of possibility of how to acquire the most suitable strategies for the each agent in multi-agent systems under the dynamically changing environment just like a soccer game as an employed example. The expected and acquired suitable strategies could be respected as the coordinated motions within agents and then evolve by themselves through the experiences of playing games. Here based on the stochastic learning automata, the agent MATEUS is proposed and the soccer games are executed by MATEUS and the result shows that an expected co-evolutionary strategies is emergently generated.

強化学習を用いたサッカーゲームモデル

○久保正男（北大院）　嘉数侑昇（北大）
Masao KUBO*　Yukinori KAKAZU**
*Division of Information Engineering
Hokkaido University
**Department of Precision Engineering,
Hokkaido University
TEL&FAX:(+81)-11-736-3818
e-mail: kubo@hupe.hokudai.ac.jp

1, アプローチ

はじめに，ここで扱うサッカーの形式的記述を行う．

$A_j^i(t) = \{ r_j^i(t), k_j^i(t), \vec{v}_j^i(t), \vec{p}_j^i(t), Unk_j^i(t) \}$ (1)

$B(t) = \{ \vec{v}(t), \vec{p}(t) \}$ (2)

$G^i \subseteq F_x \times F_y$ (3)

$F_x, F_y \in \mathfrak{R}$ (4)

ここで，iはチーム，jはプレイヤーの番号，$r_j^i(t)$，$k_j^i(t)$はチームiのj番目のプレイヤーの走力，キック力，$Unk_j^i(t)$はチームiのj番目のプレイヤーの声や姿勢や表情等の意思伝達，$\vec{v}$は移動ベクトル，$\vec{p}$はフィールド上での座標，G^iはチームiのゴール，フィールドのサイズF_x, F_yである．

1,1 コンベンショナルアプローチ

従来の評価関数を用いた手の探索による手法は以下のように記述できる．

$f^i \mid W(t) \Rightarrow \cup_{t=t..\Delta t} \cup_{i,j} A_j^i(t)$ (5)

$W(t) = \cup_{i,j} A_j^i(t) \cup B(t) \cup_i G^i$ (6)

f^iはチームiの行動出力関数，$W(t)$は環境の記述である．f^iを用いて手を探索する為には，フィールドのサイズF_x, F_yに応じて行動出力関数f^iを変える必要がある．また$W(t)$のプレイヤーの能力$r_j^i(t), k_j^i(t)$は正確に知ることは出来ず，仮定値を用いなければならない．従って，チェスや将棋等に用いられる評価関数を用いて手を探索し，各プレイヤーに行動を割り当てる方法は困難である．

1,2 マルチエージェントアプローチ

マルチエージェントシステムは個々のプレイヤーをさすエージェントが個々に観測を行い，行動$A_j^i(t+\Delta t)$を自己の判断f_j^iに基づいて生成するシステムである．つぎのように形式化できる．

$obsev_j^i(\cup_i W_j^i(t))$ (7)

$f_j^i \mid obsev_j^i(\cup_i W_j^i(t)) \Rightarrow \cup_{t=t..\Delta t} \cup_j A_j^i(t)$ (8)

ここで$obsev_j^i(\cup_i W_j^i(t))$はチーム$i$の$j$番目のプレイヤーの観測結果，$f_j^i$はチーム$i$の$j$番目のプレイヤーの行動出力関数，である．従って，MASには明確にはスーパーバイザーが存在せず，環境の微妙な変化への適応性，即応性に優れている．しかし，根本的に一つの大きな欠陥がある．それは個々のエージェントが自分勝手な行動を生成し，システム全体として適切でない危険があることである．逆に言えば，全体としての適切な行動（協調動作）の環境への適切さがシステムの能力を左右する．これを解決する為に，一般に２つの手法が提案されている．

1,3 協調動作

$Ge^i = \{ Ge_l^i(t) \}$ (9)

$Ge_l^i(t) \mid \psi_l^i \Rightarrow \cup A_j^i(t+\Delta t)$ (10)

$Ge_l^i(t)(\psi_l^i) >> f_{jk}^i(obsev_j^i(\cup_i W_j^i(t)))$ (11)

$\psi_l^i \subseteq \cup_{i,j} A_j^i(t) \cup obsev_j^i(\cup_i W_j^i(t))$ (12)

ここでGe^iはチームiの問題に対する理想的な協調動作規則集合，$Ge_l^i(t)$はチームiがもつ理想的な協調動作規則の一つである．または優先度を表す．(11)はチームがもつ協調動作規則によって，各エージェントの行動を変更できることを示している．そこで，先験的に$Ge_l^i(t)$を与えて，協調動作を生む手法が提案されている．しかし$Ge_l^i(t)$が複雑になればなるほど構築が困難である．

２つめは学習によって$Ge_l^i(t)$を獲得する手法である．つまり，ある時点までの行動履歴$\cup_j f_j^i(obsev_j^i(\cup_i W_j^i(t)))$が適切な$Ge_l^i(t)$を満たすことを学習目標とする．この手法は対象への速やかな解決は望めないが，対象の問題変化やシステム内の変化，例えばエージェント数の増減等への適応能力に優れている．ここでは強いサッカーチームを作る為に，後者のMASにおけるf_j^iを，オンライン強化学習手法の一種である確率的学習オートマトン(SLA)を用いて構築する．

1,4 エージェント"MATEUS"

提案するエージェント"MATEUS"は一般的なリアクティブプランニング手法と同様，基本的には有限個の要素からなる行動出力関数集合f_j^iから，環境の観測結果に対して適切な行動出力関数ac_{jk}^iを選択する．

$obsev_j^i(\cup_i W_j^i(t))$ (13)

$f_j^i = \{ ac_{jk}^i \}$ (14)

$ac_{jk}^i \mid obsev_j^i(\cup_i W_j^i(t)) \Rightarrow \cup_{t=t,..,\Delta t} A_j^i(t+l)$ (15)

ここで$obsev_j^i(\cup_i W_j^i(t))$はチーム$i$に$j$番目のエージェントの観測結果，$ac_{jk}^i$はチーム$i$に$j$番目のエージェントのもつ$k$番目の行動出力関数である．通常，行動出力関数$ac_{jk}^i$はデザイナーが先験的に与えるが，$ac_{jk}^i$は，はじめに述べたようにチーム内で個性的でかつ問題解決に直結していることが望ましい．一般に味方の能力や相手チームによって変化を伴うので，これを先験的に実現することは容易ではない．したがって必要なものは様々なac_{jk}^iを生成する機構であり，ここでは，オン，オフライン両方で行動集合を生成，変更し続ける機構，Action Slot，を提案する．

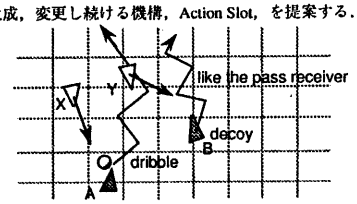

Figure 1: The decoy.

ここでサッカーを扱う理由は，サッカーゲームがもつ多様性，達成可能な$Ge_l^i(t)$が各時点で多数存在することである．例えばFigure1に示すようにエージェントBはパスレシーバーとデコイの役割のいずれも実行が可能であるが，その選択はAの行動に依存する．恐らくサッカーでは，プレイヤーの位置や速度$r_j^i(t), k_j^i(t)$に関する観測結果からだけでは，チームが満たすべき$Ge_l^i(t)$が特定できない故に，サイン等のリアルタイムでの意思伝達が行われる事が予想できる．従って各エージェントはUnk_j^iを観測し，チームが満たそうとしている$Ge_l^i(t)$を明確に察知しac_{jk}^iを選択する必要があるが，それにはUnk_j^iを具現化する必要がある．そこでMASへの緻密な協調動作生成要求に答える為に，Unk_j^iをあらわす意思伝達行動，MARKER，を与え，観測情報をふやし，協調動作候補を限定する事を試みる．MARKERはチームメイト間でのみ観測可能なフェロモンライクな物質で，すべてのフィールド上のすべてのオブジェクトに付着する．これを付着をエージェントの行動として扱い，相互に観測することにより，意思伝達を実現する．これまで意思伝達は受信者を特定し，何らかの強制を伴ったものが多く，明確に行動として扱ったものはすくなかった．

以下の節では，まずエージェント"MATEUS"を提案した後，簡単な計算機実験を行い，系の振る舞いに言及する．

2, 確率的学習オートマトンの導入

では優れたプレイヤーは，どのような情報処理を限られた時間内でおこなっているのだろうか？ここではFigure2に示すような環境観測に対する反射行動に着目する．まずはじめに核となる観測結果$obsev_j^i(\cup_i W_j^i(t))$と行動出力関数$ac_{jk}^i$のマッチングを獲得する機構を考えなければならない．

近年，学習による問題解決が盛んに研究され，多くの上記の様なマッピングを獲得する機構が提案されている．クラシファイアーシステム(C.S.)やQ-ラーニングがその手法であるが，そのパラメータ適性幅の広さや表

現の容易性に優れた確率的学習オートマトン(SLA)を用いることにする．

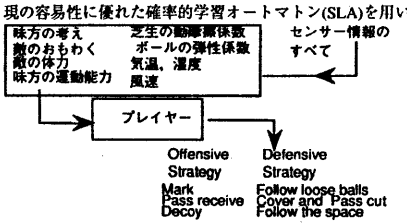

Figure 2:Ideal Player.

2.1 確率的学習オートマトン

SLAは一般に，入力状態集合$\tilde{\phi}$，出力集合$\tilde{\alpha}$，強化信号$\tilde{\beta}$，出力選択確率ベクトル集合$\tilde{P}$と評価アルゴリズム$\mathcal{A}$から構成されている．

$$SLA = \left\{ \tilde{\phi}, \tilde{\alpha}, \tilde{\beta}, \tilde{P}, \mathcal{A} \right\} \quad (16)$$

$$\tilde{\phi} = \left\{ \phi_1, \phi_2, \cdots, \phi_m \right\} \quad (17)$$

$$\tilde{P} = \left\{ \tilde{P}_1, \tilde{P}_2, \cdots, \tilde{P}_m \right\} \quad (18)$$

$$\tilde{P}_i = \left\{ P_{i1}, P_{i2}, \cdots, P_{in} \right\} \quad (19)$$

$$P_{ij} = Pr(j \mid i) \quad (20)$$

$$\tilde{\alpha} = \left\{ \alpha_1, \alpha_2, \cdots, \alpha_n \right\} \quad (21)$$

$$\tilde{\beta} = \left\{ 1,0 \right\} \quad (22)$$

ここで，mは入力状態数，nは出力数，P_{ij} は入力状態ϕ_iのとき出力α_jを行う条件付き確率を示す．従って，任意の入力状態ϕ_iにおいて，出力選択確率ベクトル集合$\tilde{P}$は条件(18)を満たす．

$$\sum_{j=1}^{n} P_{ij} = \sum_{j=1}^{n} Pr(j \mid i) = 1.0 \quad (23)$$

環境の観測結果が入力状態集合のϕ_iに含まれるならば，出力される行動は出力選択確率ベクトル$\tilde{P}_i$に従って，確率的に$\tilde{\alpha}$の中から選択される．

一方，学習は出力選択確率ベクトルの更新によって行われる．環境に出力した行動が，先験的な評価アルゴリズム$\mathcal{A}$に従って，適切ならば，強化信号1がSLAに送られ，その入力状態集合と出力集合の結合が，出力選択確率ベクトルを高めることによって，強められる．逆に出力行動が不適切であったなら（強化信号=0），結合は弱められる．たとえば，入力状態ϕ_iで，出力α_jをおこなった時に，強化信号1がSLAに与えられたならば，

$$P_{ij} \Leftarrow P_{ij} + reward \cdot (1.0 - P_{ij}) \quad (24)$$

$$P_{ikj} \Leftarrow reward \cdot P_{ikj} \quad (25)$$

を用いて，線形に更新する．ここで，kはj以外の出力を表し，rewardは更新量を表す．

2.2, 問題点

実用にあたって，一般に確率的学習オートマトンには２つの大きな課題がある．一つは観測結果を表す入力状態集合$\tilde{\phi}$の要素数mが大きくなるに従って，学習時間がのびる．式18~20に示すように，マッピングの獲得は各入力状態ϕ_iについて行うので，学習に時間がかかることは明白である．

２つめの課題は行動集合$\tilde{\alpha}$の要素数が増えると学習の効果が現われなくなることである．(24)(25)にしたがって，マッピングを表す確率を更新するが，行動数が増えると，適切な行動とそうでないものとの差が相対的に小さくなる．確率差がなくなるとパラメータの影響によって，適切な確率を維持できなくなるのである．

残念なことにサッカーはFigure2にしめすように，エージェントの観測対象は，エージェントの位置や速度$\vec{v}_j^i(t)\,\vec{p}_j^i(t)$といった連続値であり，莫大な数からなる$\tilde{\phi}$を用意しなければならない．またバラエティーに富んだ協調動作を生成するためには十分な行動出力関数ac_{jk}^iを用意しなければならい．このようなことから，$\vec{v}_j^i(t)\,\vec{p}_j^i(t)$をそのままSLAの入力状態として用いることはできず，さらに行動出力集合に至っては，多様な協調動作の生成という点から，役割と一対一対応する程度のテンプレートで構成することさえ困難である．

従って1) SLAへの入力の軽減 2) バラエティーに富んだ行動出力関数の表現という２つの課題を解決しなければならない．

3, 行動出力関数表現(Introduce CARD representation)

例えばシュートとパスを例にとり考えてみる．Figure3はAがゴールに向かってボールを蹴った結果，直接ゴールした場合と味方のエージェントに当たってゴールしたときの様子である．この２者の差はある一つの目標地点生成プロセスを異なる環境へ適用した結果と考えることはできないか．つまり，微妙な環境の違いを反映して，目標地点を生成するプロセスを用意することで，Figure2で示したような，複数の行動出力関数の代わりを果たすことが期待できる．そこで以下のような機構を提案する．

$$\widetilde{ACS}\,obsev_j^i(\cup_t W_j^i(t)) \Rightarrow \cup_{l=t,\cdots,\Delta t} A_j^i(t+l) \quad (26)$$

$$\widetilde{acs}_j^i \subset \widetilde{ACS} \quad (27)$$

$$\widetilde{acs}_j^i = \{\, acs_{jk}^i \,\} \quad (28)$$

$$ac_{jk}^i = \widetilde{acs}_{kj}^i \subset \widetilde{acs}_j^i \quad (29)$$

$$ac_{jk}^i(obsev_j^i(\cup_t W_j^i(t))) = \quad (30)$$

$$acs_{j1}^i(acs_{j2}^i(\cdots acs_{jnk}^i(obsev_j^i(\cup_t W_j^i(t)))\cdots))$$

$$\text{where} \quad nk = \left| \widetilde{acs}_{kj}^i \right|$$

ここで，$\widetilde{acs}_j^i$はチームiのエージェントjが持つ生成プロセス集合，acs_{jk}^iはその要素であるもっともプリミティブ生成プロセスである．また行動出力関数ac_{jk}^iと生成プロセス集合$\widetilde{acs}_{kj}^i$が等しいことを式(30)ように定義する．このもっともプリミティブな生成プロセスを以後カードと呼ぶ．

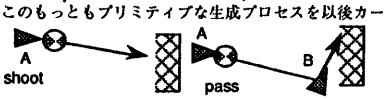

Figure 3: Shoot & Pass.

3.1, CARD

ここではカードacs_{jk}^iを用いて行動出力関数ac_{jk}^iを表現する．カードacs_{jk}^iはフィールド上のオブジェクトに関する制約条件であり，制約を満たすオブジェクトに候補値を割り当てる．

はじめに，エージェントがフィールド上の物体，ボールやゴール，プレイヤー等を認識できることを前提し，オブジェクトを定義する．

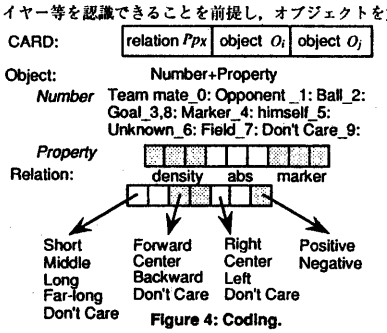

Figure 4: Coding.

次にCARDの構成を示す．CARDacs_{jk}^iは二つのオブジェクト$O_{Xi}O_{Xj}$と二者間関係Ppxという三項によってなり，オブジェクトO_iに対して関係Ppxを満たす$O_c \in O_{Xj}$を指し示す．条件を満たすO_cには各acs_{jk}^iの有効性をあらわす候補値$effe_{jk}^i$を割り当てる．当然，一つのacs_{jk}^iでac_{jk}^iを表すことは稀であり，ここではCARDの集合をCARD SET$\widetilde{acs}_{kj}^i$とよび，$acs_{jk} \in \widetilde{acs}_{kj}^i$をその効果順に適用し，候補値を足し合わせた後，しきい値処理を通して目標地点を特定する．もちろん，ディフェンス時とオフェンス時では明らかに適応すべきCARD SETは異なることが予想できる．そこで戦況に応じた$\widetilde{acs}_{kj}^i$を複数用意することにより，f_j^iを構成する．

CARDを用いることにより，一つの$\widetilde{acs}_{kj}^i$によって様々な複数のac_{jk}^iを生成することができ，SLAは適切な$\widetilde{acs}_{kj}^i$を選択することになるので，$\tilde{\alpha}$の数を減らす事が可能である．さらに思わぬ付加的な効果を得ることができる．つまりこのビット列からなる行動表現は，確率的なデフォルト階層を形成するので，これまで行動出力関数を選択する際に必要であった情報，すなわち$\vec{v}_j^i(t)\,\vec{p}_j^i(t)$，$\vec{v}(t)\,\vec{p}(t)$のほとんどをSLAの入力状態$\tilde{\phi}$から除く事ができる．従ってSLAの入力状態$\tilde{\phi}$は適切なCARD SETを選択する為の情報のみから構成できる．

3.2, MARKER

CARDを導入することによって，当初のSLAに関する課題を解決できた．そこで前記のパスシチュエーションを例にとり検証をおこなってみる(Figure1)．黒エージェントAがドリブルを行っていて，前方から二人の白エージェントがボールを奪いにきたとき，まるでBがパスを受けるかのごとく振る舞うことで，白エージェントYを引きつけることができる．こ

れは一般にデコイと呼ばれる協調動作であるが，実際にBはパスレシーバーとなっても構わない．サッカーではこのような事態はひんぱんに発生する．つまりプレイヤーやボールの位置や速度を観測するだけでは，協調動作が十分に限定できず，さらにプレイヤー間で選択した協調動作が一致しなければよい結果が得られない事態である．一般にこのような問題解法が常に膨大な量存在する問題は多様性問題と呼ばれているが，厳密な協調関係がマルチエージェントシステムに要求される．そこで多様性を解決するアプローチとして，候補が限定できない理由が情報の不足にあると考え，協調動作を限定する為にチーム固有の情報を付加することにする．つまり，チームでフィールドの見方を積極的に与えるわけである．

そこでここではより綿密な協調関係を生成獲得するために合図に対応する，MARKERを導入する．実際にプレイヤーたちはサインを送りあい意思疎通$Unk_j^i(t)$を行っているが，これをコンピュータで扱うには何らかの解釈が必要である．ここでは行動としてインクリメントし，移動やキックと同様にCARDを用いてフィールド上のオブジェクトに置かれるフェロモンライクな揮発性オブジェクトとする．例えばXがドリブルをしていて，黒チームがボールを奪おうとしている場合を考えてみる(Figure5)．恐らくAがXのボールにタックルに向かうと，XはY,Zのどちらかにパスを送ることが予想できる．そこでAはXに向かう前にY,Zの近傍にB,Cが向かう事を期待して，MARKERを置く．MARKERはチーム固有であるが，何の制約も持たないオブジェクトであるから，B,CがMARKERを適切ではあると判断して，MARKERへ向かったならばシステマティックな協調動作が実現できる．

上例からもわかるように，MARKER行動は自己の移動やキックといった行動と密接な関係がある．そこで，移動地点決定用CARD，移動地点用MARKERCARD，キック地点決定用CARD，キック地点用MARKERCARDと有効性Efをセットとして扱い，以下ではSLOTと呼ぶ．従ってCARD SETはSLOT SETに名を変える．

MARKERを導入したことで，CARDの対象はMARKERを含めたオブジェクトであり，速度や位置$\vec{v}_j^i(t),\vec{p}_j^i(t)$だけでなく，チーム内での意思伝達も観測対象となり，綿密な協調動作の獲得が期待できる．

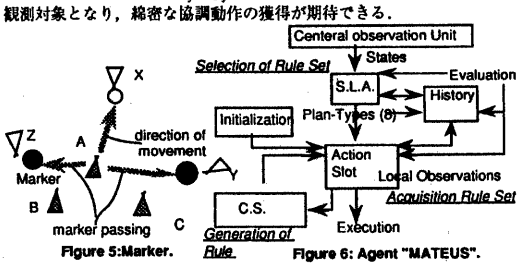

Figure 5:Marker.

Figure 6: Agent "MATEUS".

4, AGENT "MATEUS"

Figure6に本エージェント"MATEUS"の構成を示す．エージェント"MATEUS"はエージェント自身とボールとゴールとの位置関係を観測しSLAに入力する．SLAでは観測結果に適切なSLOT SETを選択する．SLOT SETでは各SLOTの有効性Efに従って確率的にSLOTを適用してゆきしきい値を越えた時点でMARKER，移動及び実行可能ならばキックを行う．

4.1, 学習

ここではその一例として，ボールを職る権利を獲得した際とゴールを決めた時にβ=1，ボールを奪われた時にβ=0が評価信号として，そのエージェントに与えられる．信号を得たエージェントはMARKERを介して，評価をチームメートに伝播する．つまりLearning Tour1) 自分が行動決定する為に使用したMARKER，Learning Tour2) 自分が発したMARKERへ移動したエージェントの行動に評価信号を時間に関する減衰定数k_tをかけた評価を伝える．評価を受け取ったエージェントは同等にSLAと用いたSLOTを線形的に$L_{R\text{-}P}$を用いて更新する．この作業を可能な限り実行する．さらに数ステップごとにSLOT間で，クロスオーバー及びミューテーションを有効性Efに基づいて実行する．

5, 結言

Figure7,8に示すのは現実的な設定下で獲得した協調動作である．センタリングやゲーム環境で頻繁に発生するスローインからの展開は極めて初期の段階で獲得されるが，複雑な協調動作の生成は乱数の影響を強く受けることがわかった．またフォワードとバックス的な役割は発生はするが，それがシステムとしての効率を反映しているとは断言できない．5400ステップ中のほとんどでいらいらさせられるといっても過言ではない(Figure9)．その理由としてカードの破壊が挙げられる．ディフェンス下では不適切な評価信号が頻繁に送られ間違った学習が進んでしまうのである．

強化学習手法を用いてサッカーでの協調動作獲得を行った．さまざまな協調動作を先験的な知識を用いず生成するために，ビット表現を行い，SLAを用いる際に発生する状態数の増加や行動集合の要素数の増加に関する欠点を回避した．最後に簡単な計算機実験をおこない協調動作の生成を確認した．さらに系の不安定性を実験を用いて示し，十分な解析が必要であることが挙げられる

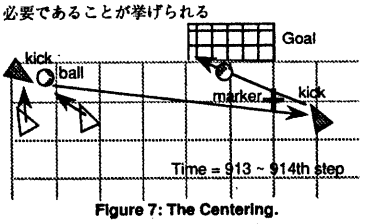

Figure 7: The Centering.

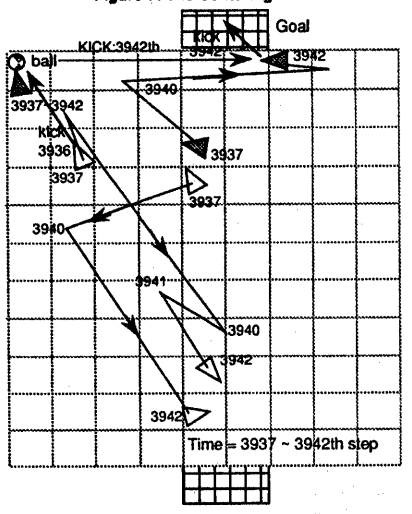

Figure 8: The soft-ball-handling.

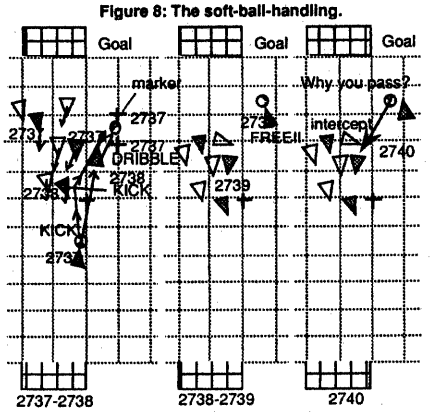

Figure 9: The miss coordinated motions.
"Why you pass?"