National Conference with International Participation
ENGINEERING MECHANICS 2003
Svratka, Czech Republic, May 12 – 15, 2003
paper no. 193

WALKING GAIT OF FOUR-LEGGED ROBOT OBTAINED THROUGH Q-LEARNING

T. Březina[1], P. Houška[2], V. Singule[3], P. Sedlák[4]

Summary: *The possible method of walking policy obtaining of four-legged robot through Q-learning is discussed in the contribution. Q-learning is implemented using architecture represented by nondeterministic state machine that defines both possible discrete states and admissible transitions between them. Discrete state is designed as indicators vector of goals achievement by single simultaneously activated instances of two basic controllers. Only simultaneous activations that guarantee static stability of robot are admissible even in the case when single activations cold not achieve its goals. The controllers attempt to achieve its goals using on-line minimization process. Q-learning sequentially improves an estimation of future benefit from usage of admissible simultaneous activations in single discrete states. Walking policy is generated through activations with the highest estimation of future benefit.*

1 Úvod

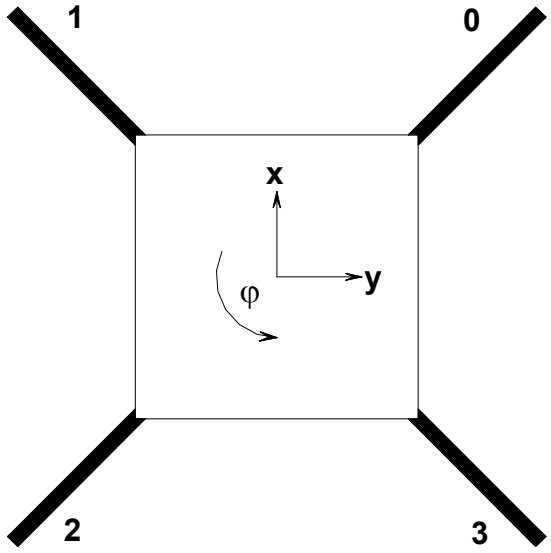
Významným cílem návrhu kráčivých robotů je realizace autonomní soustavy, která bude schopna pohybu v neznámém prostředí. Jednou z možností dosažení potřebné adaptivity řídicího systému bez potřeby modelování složitých nebo nepředvídatelných případů chování systému je využití strojového učení. Z různých paradigmat učení je velmi atraktivní Q-učení. To zaručuje, že dostatečně velkém (teoreticky nekonečném) počtu průchodů procedury učení všemi stavy prostředí a všemi stavy robotu za použití všech akcí robotu teoreticky zaručuje, že robot bude na určitý stav reagovat v jistém smyslu optimální akcí (strategií chování). Proto má pro efektivní a úspěšné použití Q-učení zásadní význam vhodná definice a diskretizace stavového prostoru prostředí i stavového prostoru robotu a rovněž vhodná volba množiny akcí. Neméně významné je, aby jak množina stavů, tak množina akcí byly co nejmenší.

Úloha je studována s použitím simulačního modelu kinematickém řízení čtyřnohého robotu v rovinném terénu s konstantní výškou těla robotu nad terénem. Cílem je dosáhnout toho, aby si robot sám v tomto terénu vyvinul optimální způsob chůze.

[1] Tomáš Březina, RNDr., Ing., CSc., VUT v Brně, FSI ÚAI, Technická 2, 61669 Brno, ČR
e-mail: brezina@uai.fme.vutbr.cz
[2] Pavel Houška, Ing., VUT v Brně, FSI ÚMT, Technická 2, 61669 Brno, ČR,
e-mail: houska@umt.fme.vutbr.cz
[3] Vladislav Singule, Doc., Ing., CSc., ÚVSSaR FSI VUT Brno, Technická 2, 616 69 Brno, ČR
e-mail: singule@zam.fme.vutbr.cz
[4] Petr Sedlák, VUT v Brně, FSI ÚMT, Technická 2, 61669 Brno, ČR,
e-mail: psedlak@volny.cz

2 Použitý přístup

V příspěvku jsou popsány první zkušenosti s diskretizací spojitého stavového prostoru čtyřnohého robotu prostřednictvím simultánních kompozic chování. Inspirací byly práce Hubera a Coelho [2,3]. Jak bylo dříve popsáno v [1], jsou simultánní kompozice chování získávány simultánními aktivacemi instancí dvou velmi jednoduchých základních řídicích členů Φ_1 a Φ_2. Instancí základního řídicího členu $\Phi_i \frac{\sigma}{\tau}$ se rozumí základní řídicí člen Φ_i, který má přiřazenu konkrétní množinu vstupních zdrojů (senzorů) $\underline{\sigma}$ a výstupních zdrojů (aktuátorů) $\underline{\tau}$ (viz. obr. 1). Základní řídicí člen konfigurace došlapu Φ_1 je definován cílem dosáhnout změnou polohy bodu došlapu jedné ze tří zvolených noh staticky stabilního postavení robotu na těchto třech nohách (dosažení trojúhelníku statické stability). Základní řídicí člen Φ_2 konfigurace noh je definován cílem dosáhnout změnou natočení těla robotu takových poloh aktuátorů noh, které jsou co nejblíže svým středním polohám (což je chápáno jako dosažení kinematicky optimálního postavení).



Obr. 1. Specifikace zdrojů robotu

Diskrétní stav robotu popisuje, ve kterých trojúhelnících stability (viz. obr. 2) se aktuálně nachází těžiště robotu a dále to, zda robot zaujímá kinematicky optimální postavení. Je tak představován pěti logickými hodnotami, tj. $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$.

Diskrétní stav souvisí s dosahováním cílů aktivovanými instancemi následovně:

$$x_1 \leftarrow \Phi_1 \frac{1,2,3}{*},\ x_2 \leftarrow \Phi_1 \frac{0,2,3}{*},\ x_3 \leftarrow \Phi_1 \frac{0,1,3}{*},\ x_4 \leftarrow \Phi_1 \frac{0,1,2}{*},\ x_5 \leftarrow \Phi_2 \frac{0,1,2,3}{x,y,\varphi},$$

kde $0, 1, 2, 3$ označují body došlapu noh robotu, x, y, φ polohu a orientaci těžiště a $*$ zastupuje koncový bod došlapu libovolné nohy, indikovaný sensory příslušné instance. Operátor přiřazení $\leftarrow$ zobrazuje výsledek činnosti instance základního řídicího členu (1 - bylo dosaženo jeho cíle, 0 - nebylo dosaženo jeho cíle) se složkou stavu robotu. Poznamenejme, že aktivace instance může ovlivnit hodnotu nejenom „svoji“ složku stavu vektoru, ale i složky další. Např. aktivace instance $\Phi_1 \frac{1,2,3}{3}$: tím, že mění polohu bodu došlapu nohy 3 robotu, může změnit výsledek dosažení cílů všech instancí, které obsahují tuto nohu ve svých vstupních zdrojích, protože tím změní navíc jednak polohu tří trojůhelníků stability a jednak postavení celého robotu.

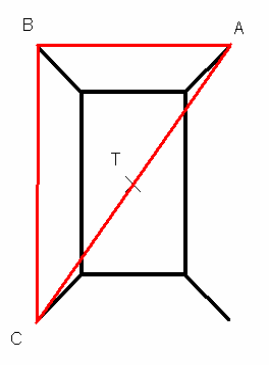
K popisu interakcí základních instancí bylo použito operátoru „podmíněnosti“ „$\triangleleft$“. Operátor omezuje řídicí proces realizovaný podřízenou základní instancí tak, aby nebylo

dotčeno dosažení cíle nadřízenou základní instancí. Zápis $\Phi_1 \frac{0,1,2}{1} \lhd \Phi_1 \frac{1,2,3}{1}$ znamená požadavek dosažení stabilního postavení v trojúhelnících $0,1,2$ a současně $1,2,3$ pohybem nohy 1, přičemž instance $\Phi_1 \frac{1,2,3}{1}$ je nadřízenou instanci $\Phi_1 \frac{0,1,2}{1}$ (podřízená instance nesmí porušovat dosažení cíle nadřízenou instancí).

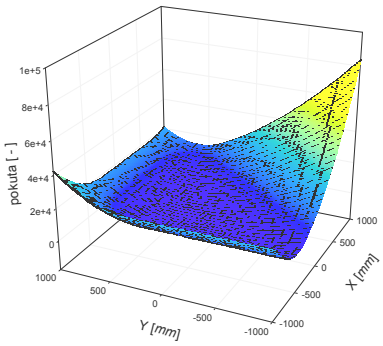
Množina přípustných stavů nesmí obsahovat stav, který by znamenal nestabilní konfiguraci robotu (např. nedosažení cíle v žádném trojúhelníku stability), a/nebo je mechanicky nedosažitelný (např. současné dosažení cílů ve dvou protějších trojúhelnících stability). Tyto podmínky apriorně omezují celkový počet stavů, které robot může zaujmout. Robot může pro změnu svého stavu použít pouze takovou simultánní aktivaci, která zaručuje, že bude vždy dosaženo některého z přípustných stavů, ať už aktivované instance dosáhly či nedosáhly svých cílů. Čistě formálními úvahami je rovněž omezen počet přípustných aktivací. Nemá např. smysl uvažovat simultánní aktivaci více než dvou řídicích členů Φ_1 a/nebo jednoho řídicího členu Φ_2, protože nemůže být současně dosaženo cíle základních instancí ve třech trojúhelnících statické stability robotu.

Q-učení pracuje s přípustnými stavy, akce jsou představovány přípustnými simultánními aktivacemi [1]. Místo obvyklé implementace dvojic stav/akce tabulkou je použito nedeterministického konečného automatu. Simultánní kompozice řídicích členů ohodnocují přechody automatu. Q-učení postupně zpřesňuje odhad budoucího prospěchu z použití jednotlivých přechodů, tj. z použití simultánních aktivací. Řízení je realizováno aktivacemi simultánních kompozic řídicích členů s nejvyšším odhadem budoucího prospěchu. Odhad budoucího prospěchu se děje na základě pokut odvozovaných z okamžité polohy těžiště a orientace robotu vzhledem k uzlovému bodu dráhy, kterého má robot dosáhnout.

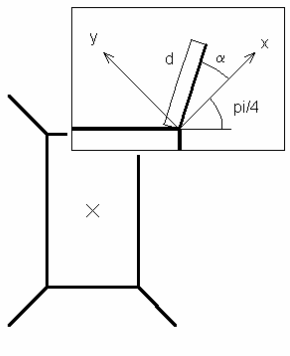
3 Implementace



Obr.2: Robot s trojúhelníkem statické stability a těžištěm



Obr.3. Pokutová funkce řídicího členu Φ_1

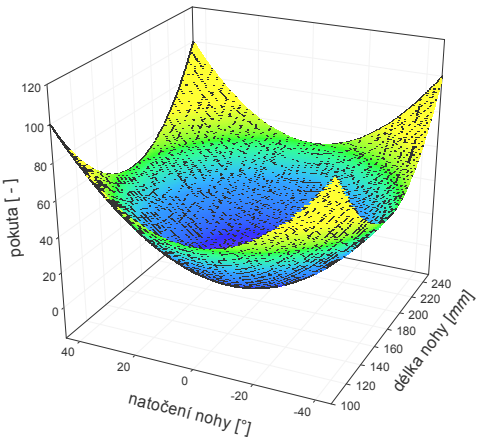


Obr.4. Délka nohy robotu d a úhlová poloha α

Řídicí členy usilují o dosažení svých cílů prostřednictvím on-line minimalizačního procesu. Byly testovány dva přístupy k realizaci tohoto procesu. V prvním přístupu odpovídá každé

instanci řídicího členu pokutová funkce, která může nabývat globálního minima pro nekonečně mnoho nezávisle proměnných. Je volena tak, aby množina nezávisle proměnných, určujících její globální minimum, odpovídala realizaci cíle dané instance řídicího členu. Průběh pokutové funkce pro Φ_1 je uveden na obr. 3, geometrie nohy plyne z obr. 4 a průběh pokutové funkce pro jednu nohu je uveden na obr. 5. Pokutová funkce Φ_2 je pak součtem pokutových funkcí všech noh.

Pokutová funkce simultánní aktivace více instancí základních řídicích členů je sestrojena jako součet pokutových funkcí jednotlivých aktivních instancí. K výsledné pokutové funkci je navíc přičtena složka, která podporuje pohyb těžiště robotu k následujícímu uzlovému bodu plánované trajektorie pohybu celého robotu. Tato složka je opět konstruována tak, že má jediné minimum v uzlovém bodě. K nalezení minima bylo použito všeobecně známých metod, a to metody nejstrmějšího sestupu a DFP.



Obr.5. Pokutová funkce délky a úhlového natočení nohy

V druhém přístupu používá výsledná pokutová funkce pouze složku, která podporuje pohyb těžiště robotu k dalšímu uzlovému bodu. Množina nezávisle proměnných, jejíž prvky představují dosažení cíle instance základního řídicího členu, je nyní vymezena nelineárními omezeními. Simultánní aktivace více instancí základních řídicích členů je realizována přidáním odpovídajících omezení do formulace minimalizační úlohy. K nalezení minima bylo použito metody sekvenčního kvadratického programování (SQP) s aktualizací Hessiánu pomocí formule BFGS.

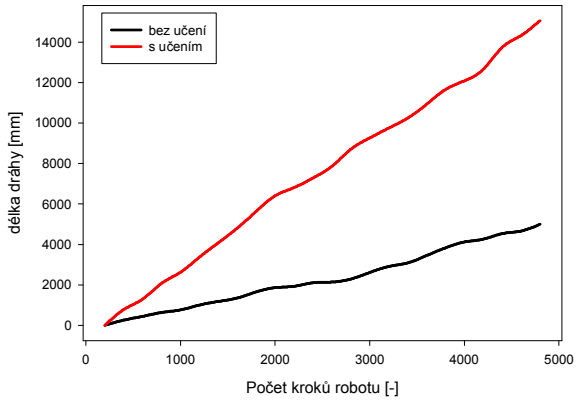
Aktivované instance řídicích členů usilují o dosažení svých cílů prostřednictvím on-line minimalizačního procesu. Pokutová funkce používá pouze složku, která podporuje pohyb těžiště robotu k dalšímu uzlovému bodu jeho dráhy.

4 První dosažené výsledky

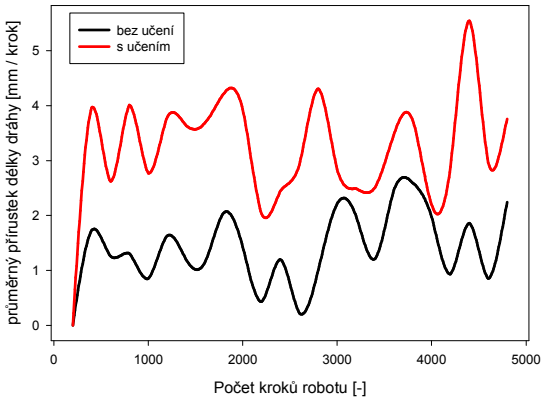
Simulačně bylo testováno chování robotu získané použitím tří minimalizačních metod, a to metody nejstrmějšího sestupu, DFP a již uvedeného SQP. Nejrychlejší byla DFP metoda s průměrnou dobou výpočtu 20 ms (AMD Duron 600MHz, 192 MB RAM), asi 2x pomalejší je metoda nejstrmějšího sestupu a SQP metoda je přibližně 15x pomalejší, než DFP metoda. Hodnotíme-li dodržení předepsaných omezení, neselhala metoda SQP ani jednou, zatímco metoda DPF selhala přibližně v 10% případů. Nejhůře se chovala metoda nejstrmějšího sestupu, která selhala asi ve 30% případů.

Chování robotu s SQP vyjadřují závislosti na obr. 6 až 9. V těchto obrázcích je vynesen vývoj délky dráhy resp. změn orientace robotu v závislosti na počtu kroků, které robot provedl během pohybu v přímém směru. V případě, že je použito učení, odpovídá každý krok

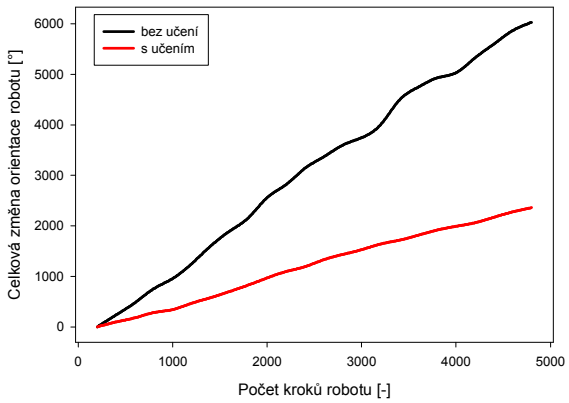
kroku Q-učení (průzkumný krok). Vynášeny jsou pro přehlednost hodnoty průměrované za každých 200 kroků.



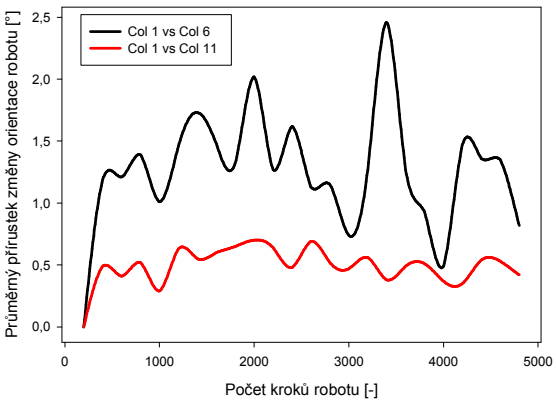
Obr. 6: Vývoj celkového přírustku délky dráhy



Obr. 7:Vývoj průměrného přírustku délky dráhy



Obr. 8: Vývoj celkové změny orientace robotu



Obr. 9: Vývoj průměrného přírustku celkové změny orientace robotu

Jak je patrno z obr. 6 a 8, robot s čistě náhodnou volbou simultánních aktivací řídících členů urazí během 4800 kroků dráhu 4500 mm při celkové změně orientace v jednotlivých krocích 6100°. Robot náhodně mění polohu a postavení noh, přičemž stále zůstává ve staticky stabilním postavení. Umožňuje-li mu to jeho aktuální konfigurace, posune střed plošiny směrem k cílovému bodu (uzlovému bodu dráhy). Při provádění „průzkumných“ kroků (kroků Q-učení) urazí během téhož počtu kroků dráhu 15000 mm při celkové změně orientace 2300°. Dosažený způsob chůze však zdaleka není optimální. Průměrné přírustky délky dráhy robotu (obr. 7) kolísají v značně širokém rozsahu. Robot dosahuje po provedení 300 kroků Q-učení přírustku délky dráhy 3,9 mm na krok, který však po dalších 300 krocích klesne na přibližně 2,6 mm na krok. Přírustky potom kolísají mezi 2 až 4,3 mm na krok, až po

provedení přibližně 4400 kroků dosáhne robot přírustku dráhy 5,5 mm na krok, který je však vzápětí následován poklesem až na 2,0 mm na krok. Robot s čistě náhodnou volbou simultánních aktivací řídících členů dosahuje přírustků délky dráhy na krok, které kolísají mezi 0,3 a 2,4 mm. Kolísání přírustků je způsobeno tím, že robot střídá fáze skutečného postupu k cílovému bodu své dráhy s fázemi, kdy provádí „průzkumné“ pohyby nohama (používá-li učení) a směrem k cílovému bodu se nepohybuje. V této fázi se může dostat i do takového postavení, že musí posunout svůj střed směrem od cílového bodu. U průměrného přírustku celkové změny orientace robotu je situace obdobná (obr. 9).

5 Závěr

Robot používající SQP se začne pohybovat směrem k uzlovému bodu své dráhy prakticky okamžitě. Je to způsobeno realizací cílů simultánních kompozic instancí řídicích členů. Jejich pokutová funkce totiž obsahuje složku, která podporuje pohyb středu plošiny robotu směrem k následujícímu uzlovému bodu. Učením je po 5000 krocího dosaženo rychlosti asi $3\times$ vyšší, než je rychlost vyvinutá bez učení (náhodnými aktivacemi řídících členů). Je však stále velmi nízká a strategie chůze, kterou si robot učením vyvine, je stále vzdálena od optimální strategie. Nízký je ale i použitý počet kroků učení. Chování robotu získané vyšším počtem průchodů procedurou učení než 5000 prozatím testováno nebylo.

První dosažené výsledky určují směr dalších prací. Ukazuje se, že značné rezervy jsou v počtu kroků Q-učení (pouhých 5000 kroků je příliš málo). Dále považujeme za slibné experimentovat zejména s nastavením parametrů procedury učení a hodnot vah pokutové funkce. Vliv konkrétní volby posilovací funkce jsme prozatím neposuzovali.

6 Poděkování

Práce byla provedena za podpory projektů MSM 262100024 „Výzkum a vývoj mechatronických soustav“, pilotního projektu ÚT AV ČR č. 52020 „Řízení kráčivého robotu s využitím metod umělé inteligence“, projektu navazujícího č. 52022 „Realizace základních řídicích členů kráčivého robotu“ a výzkumného záměru CEZ J22/98 261100009 „Netradiční metody studia komplexních a neurčitých systémů“.

7 Literatura

[1] Březina, T., Houška, P., Singule, V.: Learning Based Control System of Four-Legged Robot, Sborník nár.konf. Inženýrská mechanika 2002, Svratka, 2002, str. 9-10.
[2] Coelho, Jr. J.A., Multifingered Grasping: Grasp Reflexes and Control Context, Ph.D. Dissertation, Univ. of Massatchusetts, Amherst, 2001.
[3] Huber M., A Hybrid Architecture for Adaptive Robot Control. Ph.D. Dissertation, Univ. of Massatchusetts, Amherst, 2000.