Proceedings of the JSME Conference on
Robotics and Mechatronics
2A1-F11, May 2009

ボトムアップ視覚注視とその連続性評価によるタスク呈示からのキー動作の抽出

Examining Continuity in Bottom-Up Visual Attention Extracts Key Actions from Task Demonstration

正　長井 志江（ビーレフェルト大学）

Yukie NAGAI, CoR-Lab, Bielefeld University, Germany, yukie@techfak.uni-bielefeld.de

This paper presents a biologically-inspired model employing bottom-up visual attention for robot task learning. Although bottom-up attention enables robots to detect likely important information, discontinuity of the attention as well as its instability causes a challenge in being applied to action learning. The proposed model overcomes the problem by examining spatial and temporal continuity in low-level features for attended locations. Retina filtering and stochastic attention selection, which are integrated with saliency-based visual atteneion, facilitate the process by stabilizing the model's attention while keeping its receptiveness to a new stimulus. An experiment shows that the model can extract key actions from task demonstration.

Key Words: Bottom-up visual attention, Continuity, Learning from demonstration, Key action

1. 緒　言

本論文では，ボトムアップ視覚注視機能を拡張したロボットのタスク学習モデルを提案する．視覚的顕著性モデル [1,2] に代表されるボトムアップ視覚注視は，入力画像から色やエッジ，オプティカルフローなどの低次元の特徴量において周囲刺激と異なる刺激を顕著な領域として検出し，システムの注視対象とする．これは人間の初期視覚機能に倣ったメカニズムであり，環境に関する先見的知識を必要とせずにシステムに妥当な注視対象を与えられることから，認知発達ロボティクス研究 [3] で広く利用されている．例えば，ロボットの自律的環境探索 [4,5] や注視点からの物体およびロボット身体の学習 [6,7]，そして未知な要因の多い人間とのインタラクション [8,9] がその応用例である．

その一方で，ロボットのタスク学習はこれまですべてトップダウンに制御された視覚システムに依存してきた．他者のタスク呈示からの学習では，タスクの達成に重要な物体や呈示者の身体部位が設計者によってあらかじめ定義され，ロボットはそれらを視覚上で追従することで動作を獲得した（例えば [10]）．これは，タスク学習が動きの追従という動的な問題を扱うのに対して，ボトムアップ視覚注視は入力刺激の変化に対して敏感であり，その不安定性さが動きという空間的かつ時間的に連続な情報の獲得を困難にするということに起因している．タスクを学習する上で，ロボットは周囲の外乱を無視しつつ呈示された動きを安定に追従し，かつその一方で，周囲で新たに起こりうる動きにも適切に反応しなければならない．ロボットのタスク学習にボトムアップ視覚注視を適用することは，いかにして呈示動作からタスクに重要な情報を抽出するかという問題 [11,12] に取り組むことであり，工学的にも認知発達学的にも意義深い [13,14].

そこで本論文では，生物学的知見に基づいた三つの機能を導入することで，ロボットがボトムアップ視覚注視により，呈示されたタスクから学習の基礎となるキー動作を抽出可能であることを示す．まず第一，第二の機能は，人間の視覚を模した網膜フィルタリングと確率論的注視選択である．前者は中心窩に対して周辺視野の画像鮮明度を下げることで周囲のノイズを除去し，すでに注視されている対象への注意を強化する．これに対して，後者はフィルタリング後も顕著である周囲刺激に関して，そのような刺激はタスクに関連している可能性が高いことから，顕著性の度合いに応じた確率論的な視線変化を可能にする．つまり，これら二つの機能によって視覚注視の安定性と新規刺激に対する敏感性という相反する要求が実現される．そして第三の機能として，空間的・時間的連続性の評価を導入することで，注視対象からタスクのキー動作を獲得する．ボトムアップ視覚注視によって抽出される情報は空間的にも時間的にも断片的であるが，低次元特徴量レベルにおいて連続性を評価することで，意味のある物体および動きの抽出が可能となる．本稿では，以上の三機能によってボトムアップの情報のみからいかにタスク学習が実現されるかを実験により示す.

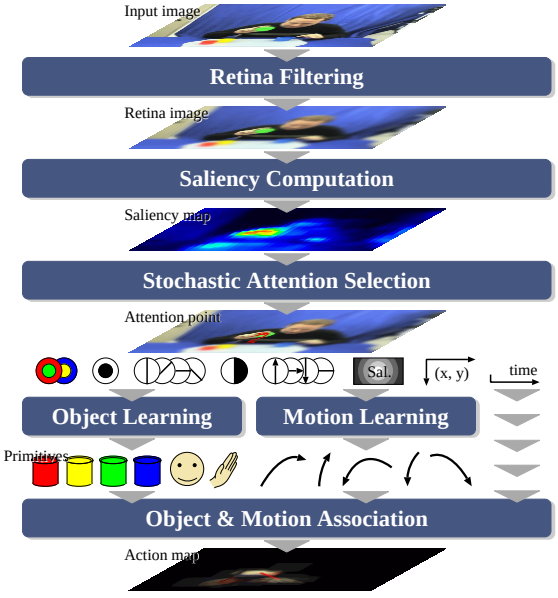


Fig. 1: A biologically-inspired model employing bottom-up visual attention for robot task learning

2. ボトムアップ視覚注視に基づくタスク学習モデル

ボトムアップ視覚注視機能を拡張したタスク学習モデルを図 1 に示す．本モデルは，タスク呈示を観察したときの視覚入力から網膜フィルタリングと視覚的顕著性の計算を経て，確率論的に注視対象を選択するモジュール群と，注視した画像領域から空間的・時間的連続性に基づき物体およびその動きを抽出し，それらを連合するモジュール群とから構成される．本モデルにより，ロボットはタスク呈示からキー動作を抽出することが可能となり，タスクを運動レベルだけではなく目的レベルで学習することが期待できる．

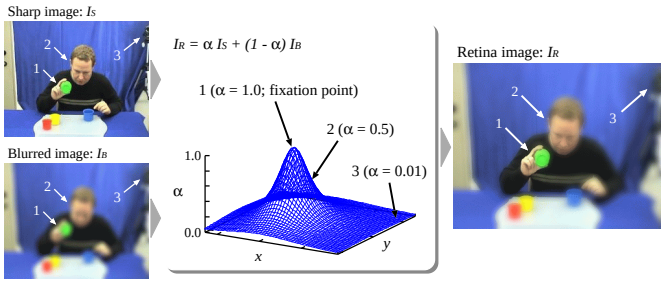
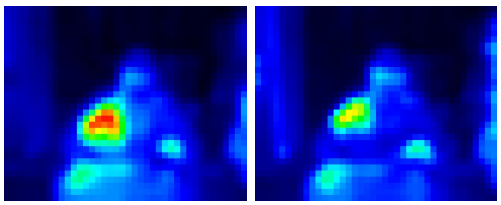


Fig. 2: Retina filtering by combining a sharp and a blurred image



(a) Derived from $\boldsymbol{I}_R$ (b) Derived from $\boldsymbol{I}_S$

Fig. 3: Saliency maps with (a) and without (b) retina filtering

2.1 網膜フィルタリング

網膜フィルタリングは人間の視覚機能を模したもので，入力画像の周辺視野の鮮明度を下げることで，中心窩の刺激の顕著性を相対的に増大させ，周辺視野からは顕著な刺激のみを検出することを可能にする．フィルタリングのメカニズムを図 2 に示す．ここでは，呈示者が赤，黄，緑色のカップを青色のカップに重ね入れるスタッキングカップタスクを呈示している．

網膜画像 $\boldsymbol{I}_R$ は，カメラから獲得される鮮明な入力画像 $\boldsymbol{I}_S$ とそれをガウスフィルタでぼかした画像 $\boldsymbol{I}_B$ の重み付き和として生成される．$\boldsymbol{x} = (x,\ y)$ を画像上の位置，$\boldsymbol{x}_F(t)$ を時刻 t の注視点とすると，画素値 $I_R(\boldsymbol{x}, t)$ は下記の式で求められる．

$$I_R(\boldsymbol{x}, t) = \alpha I_S(\boldsymbol{x}, t) + (1 - \alpha) I_B(\boldsymbol{x}, t) \tag{1}$$

$$\text{where} \quad \alpha(\boldsymbol{x}, t) = \frac{D^2}{\|\boldsymbol{x} - \boldsymbol{x}_F(t)\|^2 + D^2} \tag{2}$$

$\alpha(\boldsymbol{x}, t)$ は中心位置を $\boldsymbol{x}_F(t)$，最大値を 1.0, 直径を D とするコーシー分布関数である．これによって図 2 の例では，位置 1（注視点，$\alpha = 1.0$）において $\boldsymbol{I}_R$ は $\boldsymbol{I}_S$ と同じ鮮明さを保った画像となり，位置 2，3 ではそれぞれ位置 1 からの距離に応じた不鮮明度の画像が投影される．

2.2 視覚的顕著性の計算

次に，システムの注視位置をボトムアップに決定するため，$\boldsymbol{I}_R$ から視覚的顕著性を計算する．ここでは，Itti et al. [1,2] によって提案されたモデルを適用し，色彩，明度，エッジ，フリッカー（明度の時間差分），オプティカルフローの五種類の特徴量に関して顕著性を計算し，その総和を用いる．メカニズムの詳細は文献 [1,2] を参照されたい．

図 3 に視覚的顕著性の計算結果と，そこに現れる網膜フィルタリングの効果を示す．図 3 (a) が $\boldsymbol{I}_R$ から算出した顕著性マップ，(b) が $\boldsymbol{I}_S$ から算出したマップである．どちらにおいても，呈示者の右手位置が顕著な領域として検出されている．これは，カップを操作することによる動きと，カップや手自体がもつ固有の静的特徴に起因している．また図 3 (a) と (b) の比較から，網膜フィルタリングによって中心窩の顕著性が増幅され，反対に周辺視野のそれが減少されていることが分かる．これが，結果的に視覚注視の安定化につながる．

2.3 確率論的注視選択

システムは算出された顕著性に基づき，次時刻 $t+1$ の注視点を確率論的に決定する [15,16]．まず各画像位置 $\boldsymbol{x}$ に関して，現在の注視点 $\boldsymbol{x}_F(t)$ からの注視の遷移確率 $\phi(\boldsymbol{x}, t+1)$ を，顕著度 $s(\boldsymbol{x}, t)$ を用いて計算する．

$$\phi(\boldsymbol{x}, t+1) = \frac{\exp\left(-\beta\left(s\left(\boldsymbol{x}_F(t), t\right) - s(\boldsymbol{x}, t)\right)\right)}{\sum_{\boldsymbol{x}'} \exp\left(-\beta\left(s\left(\boldsymbol{x}_F(t), t\right) - s(\boldsymbol{x}', t)\right)\right)} \tag{3}$$

ここで β は正の定数である．つまり，顕著度が現在の注視点のそれより高いほどその画像位置は次時刻の注視点として高確率で選択され，それは顕著性の差に対して指数関数で増大する．そして，システムは $\phi(\boldsymbol{x}, t+1)$ に基づき次の注視点候補 $\boldsymbol{x}_{F'}$ を選択し，それが

$$\Delta s(\boldsymbol{x}_{F'}, t) = s(\boldsymbol{x}_{F'}, t) - s\left(\boldsymbol{x}_F(t), t\right) > 0 \tag{4}$$

の条件を満たすか，もしくは

$$p(\boldsymbol{x}_{F'}, t) = \exp\left(\Delta s(\boldsymbol{x}_{F'}, t)/T\right) \tag{5}$$

の確率を満たす場合に $\boldsymbol{x}_{F'}$ を次の注視点 $\boldsymbol{x}_F(t+1)$ として採用する．ここで，T は確率論的過程のランダムさを定義しており，この選択過程は注視点候補が上記の条件を満たすまで繰り返される．これによって，システムは顕著性の若干低い周辺刺激に対しても反応することが可能となり，新規刺激に対する敏感性が実現される．

2.4 空間的・時間的連続性の評価による物体と動きの抽出

上記の過程で選択される注視点は画像の微小領域にすぎず，必ずしも物体の全体像を捉えているとは限らない．また，各時刻で独立に注視点が決定されるため，現在注視している対象が先時刻の注視対象と同じであるという保証もない．そこで，システムは注視点の低次元特徴量に関して空間的かつ時間的な連続性を評価することで，意味のある物体とその動きとを抽出する．

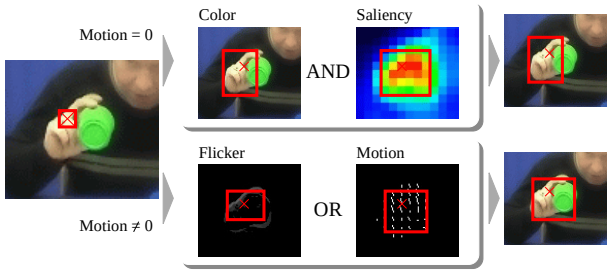
2.4.1 物体の抽出

物体抽出のメカニズムを図 4 (a) に示す．システムは注視点の運動状態に応じて，それが静止している場合（図上段）には色彩値と顕著度の類似性を評価し，運動している場合（図下段）にはフリッカー値とフロー値の類似性を評価することで物体領域を決定する．図 4 (a) の例では注視点（× 印）は呈示者の右手にあるが，動きの類似性を利用することで，右手が操作するカップも同時に検出することができる．このように一体となって運動する領域をひとつの物体として扱うことで，より安定した動きの追従が可能になる．

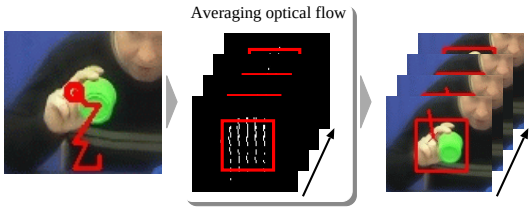
次に，抽出した物体画像を図 5 左上に示すように時間的連続性に基づいて編成する．連続する物体画像が類似の色ヒストグラムを有しかつその検出位置が十分に近い場合は，それらを同じ物体チャンクとして認識し，色ヒストグラムもしくは検出位置が大幅に変化した時点で新しいチャンクを生成する．図 5 の例では，まず呈示者が緑色のカップを操作している間に肌色と緑色から成る物体チャンクが形成され，呈示者がそれを青色のカップに挿入した時点で青色がより支配的になり，新たなチャンクが生成される．その後，呈示者が黄色のカップを操作し始めると，それはまた新たなチャンクとして認識される．

2.4.2 動きの抽出

動きの抽出も物体抽出と同様に行われる．図 4 (b) に示すように，まず注視点を空間的に拡張した物体領域からフローを検出し，



(a) Selective object extraction



(b) Generative motion detection

Fig. 4: Object and motion extraction by spatial continuity

それを各時刻ごとに平均化することでより精度の高い運動ベクトルを得る．その後，図 5 左下に示すように運動ベクトルを時間軸に沿ってならべ，それらが類似の位置および方向性を有している場合は同じチャンクとして認識する．この例では，緑色のカップを上に持ち上げたときの運動チャンク，それを青色のカップに納めるまでの右下方向の運動チャンク，そして呈示者の右手が次の黄色いカップに移るまでの二つの左方向運動チャンクが形成されている．

2.5 物体と動きの連合学習

抽出された物体と動きは，最終的にキー動作を抽出するため空間的かつ時間的連続性に基づいて連合される．図 5 にそのメカニズムを示す．前節で説明したとおり，この例では三つの物体チャンクと四つの運動チャンクがすでに生成されている．

t_{Oi} と t_{Mj} をそれぞれ i 番目の物体チャンクと j 番目の運動チャンクが生成され始めた時刻とする．二つの連続する物体チャンク間の変位 $\boldsymbol{X}(i)$ は

$$\boldsymbol{X}(i) = \boldsymbol{x}_F(t_{Oi}) - \boldsymbol{x}_F(t_{Oi-1}) \tag{6}$$

で求められ，これに対応する運動ベクトル $\boldsymbol{M}(i)$ は

$$\boldsymbol{M}(i) = \sum_{t_{Oi-1} < t_{Mj'} \leq t_{Oi}} \boldsymbol{m}(j') \tag{7}$$

として計算される．ここで，$\boldsymbol{m}(j)$ は j 番目の運動チャンクの累積運動ベクトルである．これに基づき，物体チャンクと運動チャンク間の連合は，空間的・時間的連続性の概念から，以下の条件を満たすときに確立される．

$$\arccos \frac{\boldsymbol{X}(i) \cdot \boldsymbol{M}(i)}{\|\boldsymbol{X}(i)\| \, \|\boldsymbol{M}(i)\|} < \theta \tag{8}$$

ここで，θ は閾値である．つまり，もし二つの連続する物体チャンク間の変位がその間に検出された累積運動ベクトルと一致するならば，それらは一つの意味のある動作として認識される．図 5 の例では，緑色カップの物体チャンク $O1$ と青色カップの物体チャンク $O2$ が運動チャンク $M2$ で結合され，$O2$ はその後黄色カップの物体チャンク $O3$ と $M3$，$M4$ で結合される．これが最終的に動作マップとなり，呈示タスクのキー動作（ここでは，カップを持ち上げる，カップを重ね入れる，次のカップを掴む，の三つ）を連合した地図が形成される．ここで，例えば呈示者の顔が物体チャンクとして検出されていたとしても，それは動作マップ上では他のチャンクと結合をもたないことに注意されたい．呈示者の顔が検出される位置と他の物体チャンク間の変位は，この期間に検出される運動ベクトル（主に，カップ操作から検出される手の動き）と何の関係ももたないことから，顔チャンクはタスクに関係のない物体として除外することができる．

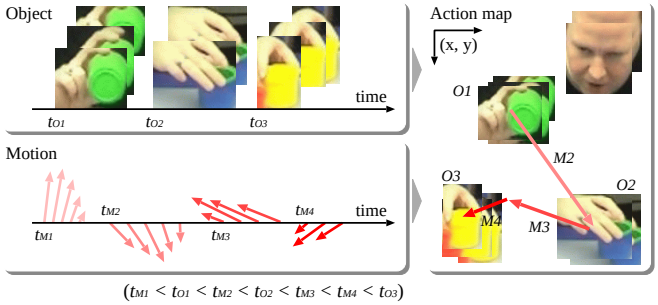


Fig. 5: Object and motion association by spatial and temporal continuity

3. 実 験

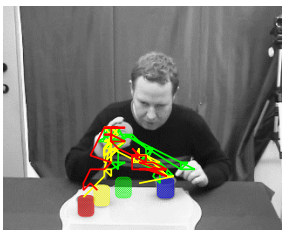
3.1 ボトムアップ視覚注視の安定性と敏感性

ボトムアップ視覚注視の安定性と新規動作に対する敏感性を評価する実験を行った．図 6 に比較実験の結果を示す．図 6 (a) が網膜フィルタリングと確率的注視選択の両方を採用したときの注視点の移動軌跡，(b) が網膜フィルタリングのみを利用し，各時刻において最も顕著な位置を注視点として選択したときの軌跡，(c) が網膜フィルタリングなしで (b) と同じアルゴリズムで注視点を選択したときの軌跡である．赤，黄，緑色の線がそれぞれの色のカップを移動させたときの注視点軌跡に対応している．

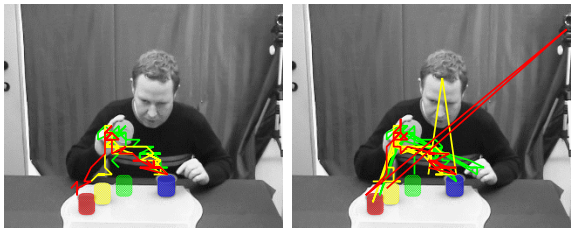
まず，図 6 (b) と (c) の比較から，網膜フィルタリングによって注視軌跡が安定化されていることが確認できる．図 6 (b) では各色のカップがその移動中に適切に追従され，タスクに無関係な場所（例えば，画像右上に映るカメラ）が一度も注視されていないことが分かる．そして，図 6 (a) と (b) を比較することにより，確率的注視選択が新規動作への敏感性を実現していることが確認できる．注目すべきは，図 6 (a) において緑色の線が呈示者の左手に到達している点である．これは，ちょうど画像に映るように呈示者が左手で青色のカップを指さしたときに，システムの注視点が呈示者の右手から左手へと移ったためで，周辺視野の動きに対しても迅速に反応できたことを示している．呈示者のこの動作は右手で操作するカップの目標位置を指示しており，タスクの目的を理解する上で重要な情報である．

3.2 空間的・時間的連続性の評価によるキー動作の抽出

次に，注視対象の連続性の評価によってキー動作がいかに抽出されるかを検証した．タスク呈示をとおして抽出された物体チャンクと運動チャンクの時間的遷移を図 7 (a) に示す．この結果から，三つのカップ操作に関する物体チャンク（上段）とそれらに対応する運動チャンク（下段，赤線）がそれぞれ適切に獲得されていることが確認できる．物体抽出では緑，黄，赤色のカップとそれらを操作する呈示者の右手が，その操作順に物体チャンクとして検出されている．そして各カップに関して，それを持ち上げ（上方向の運動ベクトル），青色のカップに挿入し（右下方向の運動ベクトル），次のカップに移る（左方向の運動ベクトル）という三つの運動チャンクがそれぞれ抽出されていることが分かる．



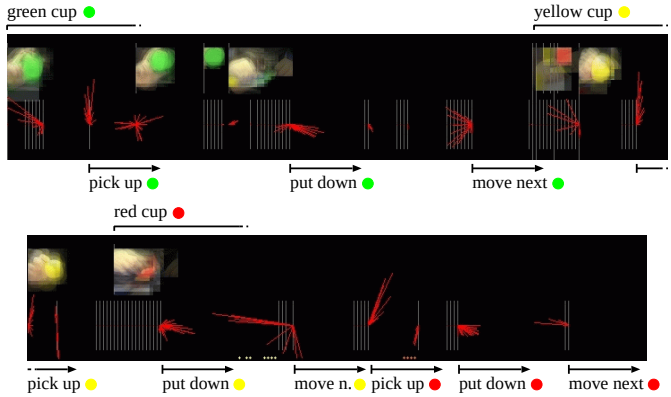
(a) Stochastic attention selection *with* retina filtering



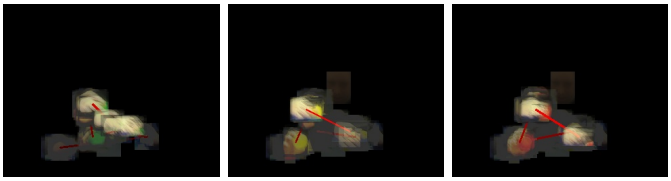
(b) Winner-take-all *with* retina filtering

(c) Winner-take-all *without* retina filtering

Fig. 6: Attention transition of proposed model (a) and two comparative models, (b) and (c)



(a) Object and motion chunks created through task demonstration



(b) Action map for moving green, yellow, and red cups

Fig. 7: Extracted key actions and their association represented in time and space

そして，これらの物体，運動チャンクを空間的・時間的連続性に基づいて連合した結果が図 7 (b) に示す動作マップである．ここでは理解を容易にするため，各色のカップごとに別々にマップを表示した．この結果から，それぞれのカップ操作において三つのキー動作を頂点とした三角形のパターンが獲得されていることが分かる．左下頂点がカップを掴む動作，上頂点がカップを持ち上げる動作，そして右下頂点がそのカップを青色カップに納めるという動作に対応している．以上の結果は，提案したモデルが呈示タスクを単なる運動軌跡としてではなく目的レベルで理解できる可能性を示唆している．

4. 結 言

本論文では，ボトムアップ視覚注視機能の拡張とそれから検出される情報の連続性の評価によって，呈示タスクからキー動作を抽出可能であることを示した．タスクの先見的知識なしに，システムが低次元の刺激特徴のみからいかにタスクに重要な情報を抽出できるかを示すことは，工学的視点からも認知発達学的視点からも非常に意義深い．本稿で提案したモデルは生物学的知見に基づいており，ロボットの自律学習の促進だけではなく，高次の認知機能が未熟な幼児のタスク学習メカニズムの解明にもつながると期待される．今後は，システムに物体や動きの学習機能を追加することで，異なるタスクの呈示からいかに汎用的なキー動作を抽出し，それを実機ロボットで再現できるかを検証していく．

文 献

[1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] L. Itti, N. Dhavale, and F. Pighin, “Realistic avatar eye and head animation using a neurobiological model of visual attention,” in *Proc. of the SPIE 48th Annual Intl. Symp. on Optical Science and Technology*, 2003, pp. 64–78.

[3] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, “Cognitive developmental robotics as a new paradigm for the design of humanoid robots,” *Robotics and Autonomous Systems*, vol. 37, pp. 185–193, 2001.

[4] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, “Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub,” in *Proc. of the IEEE Intl. Conf. on Robotics and Automation*, 2008, pp. 962–967.

[5] N. J. Butko and J. R. Movellan, “I-pomdp: An infomax model of eye movement,” in *Proc. of the IEEE Intl. Conf. on Development and Learning*, 2008.

[6] C. C. Kemp and A. Edsinger, “What can i control?: The development of visual categories for a robot’s body and the world that it influences,” in *Proc. of the 5th IEEE Intl. Conf. on Development and Learning*, 2006.

[7] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, “Visual attention by saliency leads cross-modal body representation,” in *Proc. of the 7th IEEE Intl. Conf. on Development and Learning*, 2008.

[8] L. Aryananda, “Attending to learn and learning to attend for a social robot,” in *Proc. of the 6th IEEE-RAS Intl. Conf. on Humanoid Robots*, 2006, pp. 618–623.

[9] C. Muhl and Y. Nagai, “Does disturbance discourage people from communicating with a robot?” in *Proc. of the 16th IEEE Intl. Symp. on Robot and Human Interactive Communication*, 2007, pp. 1137–1142.

[10] A. Billard and R. Siegwart, “Special issue: Robot learning from demonstration,” *Robotics and Autonomous Systems*, vol. 47, no. 2–3, 2004.

[11] C. L. Nehaniv and K. Dautenhahn, “Like me? measures of correspondence and imitation,” *Cybernetics and Systems: An International Journal*, vol. 32, pp. 11–51, 2001.

[12] C. Breazeal and B. Scassellati, “Challenges in building robots that imitate people,” in *Imitation in Animals and Artifacts*, K. Dautenhahn and C. L. Nehaniv, Eds. MIT Press, 2002, pp. 363–389.

[13] Y. Nagai, C. Muhl, and K. J. Rohlfing, “Toward designing a robot that learns actions from parental demonstrations,” in *Proc. of the 2008 IEEE Intl. Conf. on Robotics and Automation*, 2008, pp. 3545–3550.

[14] Y. Nagai and K. J. Rohlfing, “Parental action modification highlighting the goal versus the means,” in *Proc. of the IEEE 7th Intl. Conf. on Development and Learning*, 2008.

[15] D. Brockmann and T. Geisel, “The ecology of gaze shifts,” *Neurocomputing*, vol. 32–33, pp. 643–650, 2000.

[16] G. Boccignone and M. Ferraro, “Modelling gaze shift as a constrained random walk,” *Physica A*, vol. 331, no. 1, pp. 207–218, 2004.