1I3-2

マルチモーダルLDAを用いたロボットによる多様な概念の形成

The Formation of Various Concepts by Robots Using Multimodal LDA

安藤義記 Yoshiki Ando　中村友昭 Tomoaki Nakamura　長井隆行 Takayuki Nagai

電気通信大学大学院情報理工学研究科
Faculty of Informatics and Engineering, The University of Electro-Communications

In recent studies, it has been revealed that robots can form concepts and understand the meanings of words through inference. The key idea underlying these studies is “multimodal categorization” of a robot’s experience. However, previous studies considered only object categories. Our concept considered not only object categories, but also tactilecategories and color categories, which are directly connected to themodalities. In this paper, by extending multimodal latent Dirichlet allocation (MLDA), we propose the formation of various categories based on the ties with modality. We show that a robot can form various concepts based on self-obtained multimodal information.

1. はじめに

事物のカテゴリ分類は，人間の認知機能において重要な役割を果たしていることが指摘されている．人間はカテゴリを形成することで，経験した物事を全て参照することなく，必要最小限の認知的処理によってより多くの情報を得ることができる [Rosch 99]．さらに，カテゴリ分類の重要性は，経験を通して形成したカテゴリを利用した予測が可能な点にある．人間は，未知の物事に対しても様々な予測を行い，柔軟に対応している．すなわちロボットにおいても，このような経験をカテゴリ分類する能力を持つことは非常に重要であると考えられる．

これまで著者らは，自然言語処理の分野で盛んに研究されてきた統計モデルの一つである latent Dirichlet allocation (LDA) [Blei 03] をベースに，物体カテゴリを教師なしで形成する手法を提案してきた [Nakamura 08, Nakamura 09]．これらの研究では，物体の視覚や聴覚，触覚などのマルチモーダル情報を LDA によりカテゴリ分類することで，ロボットがマラカスやタンバリン，ぬいぐるみといった人間の感覚に即した物体のカテゴリ（概念）を形成できることを示した．

しかし，人が用いているカテゴリは，物体カテゴリだけでなく，モダリティに直結したカテゴリ (色カテゴリや触覚カテゴリ) 等，様々なカテゴリが存在し，複雑な構造をしている．本稿では，multimodal latent Dirichlet allocation(MLDA) を拡張し，それぞれのモダリティとの結びつきの強さを考慮した，様々なカテゴリの形成を行う．まず，物体から得られたマルチモーダル情報から，モダリティとの結びつきの強さを変化させた複数の MLDA による分類を行う．しかしながら，様々なカテゴリを形成することにより，その中には人の感覚には即さないカテゴリも多く形成されることになる．そこで，人との対話を通してカテゴリに関する単語情報を取得し，形成されたカテゴリと単語を結びつけ，単語が表すカテゴリの選択を行う．最終的に，ロボットはマルチモーダル情報により形成されるカテゴリと，それを表す単語，さらにはカテゴリとモダリティとの結びつきの強さを得ることができる．

提案手法は，確率的に様々な推論が可能であり，例えば，ロボットは物体を見ることで，その視覚情報から物体の聴覚情報や触覚情報の予測が可能となる．また，対話により概念と単語が結びつくため，ロボットが知覚した情報を単語で表現することが可能となり，人の用いる単語と結びつきの強い概念を複数所持することで，色のみに注視するなど，単一視点からの予測だけではなく，複数の視点からの様々な予測が可能となる．さらに，モダリティと単語の結びつきが獲得されるため，単語から特定のモダリティへ注意を向けること等も可能となる．

連絡先: 安藤 義記，電気通信大学大学院情報理工学研究科，東京都調布市調布ヶ丘 1-5-1，y4422@apple.ee.uec.ac.jp

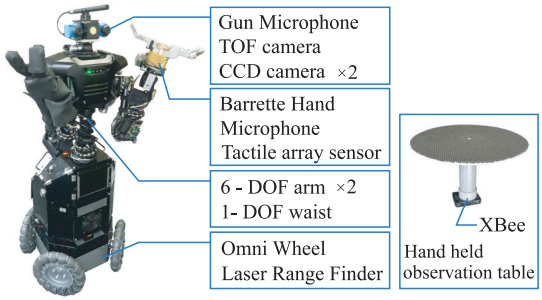


図 1: ロボットプラットフォーム

2. 提案手法

2.1 マルチモーダル情報の処理

本稿では，図 1 のロボットを用いることを想定する．ロボットは，物体を発見し自律的にマルチモーダル情報を取得する [Araki 12]．ここでは取得するマルチモーダル情報と，その処理に関して述べる．

視覚情報

まず観測した物体の画像を複数枚取得する (後述する実験では，各物体に対して 36 枚の画像を取得した)．本稿では特徴量として 128 次元の DSIFT を用い，これにより 1 枚の画像から多数の特徴ベクトルを得ることができる [Vedaldi 10]．これらの特徴ベクトルを，学習画像とは関係のない背景画像から計算した 500 の代表ベクトルを用いてベクトル量子化することで得られる 500 次元のヒストグラムを視覚情報として取り扱う．

さらに，2 つ目の視覚情報として，Lab 表色系の補色次元 a 及び b の 2 次元ヒストグラムを用いた．ビンの数はそれぞれ 5 とし，合計 25 次元のヒストグラムとした．

聴覚情報

取得した音情報は 0.2[sec] 毎のフレームに分割し，フレーム毎の特徴量に変換する．特徴量としては，音声認識でよく利用されている MFCC を用い，各フレームは 13 次元の特徴ベクトルとなる．これにより，物体から発生した音から複数の特徴ベクトルを得ることができる．この特徴ベクトルを，あらかじめ計算した 50 個の代表ベクトルによりベクトル量子化を行

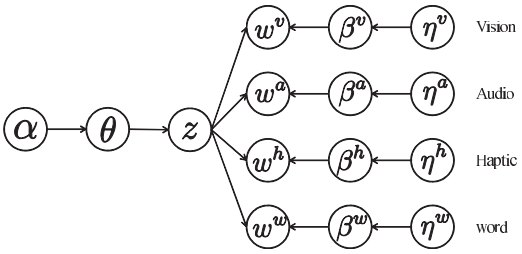


図 2: MLDA のグラフィカルモデル

い，各代表ベクトルの発生頻度を表すヒストグラムを聴覚情報として使用する．

触覚情報

触覚情報には，162 個のセンサから構成された触覚センサにより取得した時系列データを用いる．取得したデータは近似を行い，近似パラメータを各センサの特徴ベクトルとして扱う [Nakamura 10]．さらに k 平均法により予め計算した 15 の代表ベクトルを用いてベクトル量子化を行い，15 次元ヒストグラムを触覚情報として用いる．

単語情報

ロボットは，物体を観察中に人から発せられた教示発話を単語情報として利用する．教示発話のうち，物体の特徴を表す単語を選び，最終的に，単語の発生頻度ヒストグラムを単語情報として用いる．

2.2 カテゴリ分類と概念形成

本稿では，ロボットが経験することによって得るマルチモーダルな情報をカテゴリ分類して形成した各カテゴリを概念として考える．つまり概念は，特徴空間上のクラスタとして表現されており，そのクラスタを用いることで，ある一部の入力から観測されなかった次元の情報を予測することが可能となる．言語の情報も特徴空間の一部となっており，概念に基づく予測のメカニズムが語意の理解や言語表現の基盤となっている．こうした分類や予測を確率的に実現するために，次に述べるマルチモーダル LDA(MLDA) を用いる．

2.3 マルチモーダル LDA

MLDA は，LDA [Blei 03]　のマルチモーダル情報への拡張であり [Nagai 12]，図 2 のグラフィカルモデルで表される．図中の w^v, w^a, w^h, w^w は，それぞれ視覚・聴覚・触覚・単語情報を表しており，β^* をパラメータとする多項分布から生成される．また，β^* は，η^* をパラメータとするディリクレ事前分布によって決定される．z は物体のカテゴリを表しており，θ をパラメータとする多項分布から生成される．同様に，θ は α をパラメータとするディリクレ事前分布によって決定される．ここでカテゴリ分類の問題は，観測したマルチモーダル情報に基づき，モデルのパラメータを推定することに帰着される．図 2 から分かるように，MLDA は観測された情報から，観測されていない情報を確率的に推論する枠組みを提供しており，これが予測に基づく理解の基本的な仕組みとなっている．

2.4 Bag of Multimodal LDA:BoMLDA

まず，MLDA を Bag of Multimodal LDA(BoMLDA) へと拡張を行う．図 3 が提案する BoMLDA のグラフィカルモデルである．BoMLDA はモダリティへの重みやカテゴリ数を様々に変化させた MLDA の集合であり，ロボットが実際に取得したマルチモーダル情報を MLDA によりカテゴリに分類することで，様々な概念の形成を行う．w^v, w^a, w^h, w^c, w^w は，それぞれ視覚 (SIFT)・聴覚・触覚・視覚（色）・単語情報であり，β^* をパラメータとする多項分布から生成される．また，z はカテゴリを表しており，θ は z の出現確率分布を表す多項分布のパラメータである．このパラメータは，ハイパーパラメータ α により決まるディリクレ事前分布に従う．さらに，λ^v, λ^a, λ^h, λ^c は，それぞれ視覚 (SIFT)・聴覚・触覚・視覚（色）情報への重みであり，K はモデルのカテゴリ数を意味する．

2.5 Gibbs Sampling によるパラメータ推定

カテゴリ分類は，マルチモーダル情報から，図 3 内のパラメータを推定することに相当する．本稿では，パラメータ推定に Gibbs Sampling を用いる．Gibbs Sampling では，j 番目の物体のモダリティm の情報の i 番目に割り当てられるカテゴリ z_{mij} は，θ，β^* を周辺化した条件付確率

$$p(z_{mij}=k|z^{-mij},w^m,\alpha,\pi^m,\phi) \propto (\sum_{m'}\lambda^{m'}N^{-mij}_{m'kj}+\alpha)\frac{\lambda^m N^{-mij}_{mw^mk}+\pi^m}{\lambda^m N^{-mij}_{mk}+W^m\pi^m} \quad (1)$$

からサンプリングされる．ただし，W^m はモーダル情報の次元数である．N_{mw^mkj} は，j 番目の物体のモダリティm の情報が w^m となり，かつカテゴリ k が割り当てられた回数を表している．また，λ^v, λ^a, λ^h, λ^c, λ^w はそれぞれ視覚 (SIFT)・聴覚・触覚・視覚（色）情報への重みを表しており，この重みによって特定のモダリティと結びついたカテゴリを形成することが可能となる．さらに，ϕ はモデルのパラメータであり，$\phi=\{K,\lambda^v,\lambda^a,\lambda^c,\lambda^w\}$ となり，$N_{mkj}, N_{mw^mk}, N_{mk}$ は以下のように表現できる．

$$N_{mkj}=\sum_{w^m}N_{mw^mkj} \quad (2)$$

$$N_{mw^mk}=\sum_{j}N_{mw^mkj} \quad (3)$$

$$N_{mwk}=\sum_{w^m,j}N_{mw^mkj} \quad (4)$$

N_{mkj} は j 番目の物体のモダリティm の情報に，カテゴリ k が割り当てられた回数を，N_{mw^mk} はモダリティm の情報 w^m にカテゴリ k が割り当てられた回数を，N_{mk} は全ての物体のモダリティm の情報に，カテゴリ k が割り当てられた回数を表している．また，式 (1) 内の除算の添え字はその情報を除くことを意味しており，z^{-mij} は j 番目の物体のモダリティm の i 番目の情報へ割り当てられたカテゴリ z_{mij} を取り除いた残りを示している．Gibbs Sampling では, 各物体 j のモダリティ m の i 番目の情報へのカテゴリの割り当てを, 式 (1) に従いサンプリングを行う．これを繰り返すことで,N_* がある値へと収束する．最終的に，パラメータの推定値 $\hat{\beta}^m_{w^mk}, \hat{\theta}_{kj}$ は以下のようになる．

$$\hat{\beta}^m_{w^mk}=\frac{\lambda^m\hat{N}_{mw^mk}+\pi^m}{\lambda^m\hat{N}_{mk}+W^m\pi^m} \quad (5)$$

$$\hat{\theta}_{kj}=\frac{\sum_m\lambda^m\hat{N}_{mkj}+\alpha}{\sum_m\lambda^m\hat{N}_{mj}+K\alpha} \quad (6)$$

ただし，N_{mj} は，j 番目の物体のモダリティm の情報の総数，$\hat{N}_*$ は，式 (1) に従いサンプリングを繰り返したことにより収束した N_* の値である．最終的に j 番目の物体のカテゴリ z_j は以下のようになる．

$$z_j=\underset{k}{\mathrm{argmax}}\,p(z=k|w^v,w^a,w^c,w^w,\phi)=\underset{k}{\mathrm{argmax}}\,\hat{\theta}_{kj} \quad (7)$$

これら分類はモデルのパラメータ ϕ によって変化する．モダリティへの重み λ^* は，特定のモダリティとの結びつきの強さを表しており，この値によって特定のモダリティと結びついたカテゴリを形成することができる．また，分類の粒度は K に

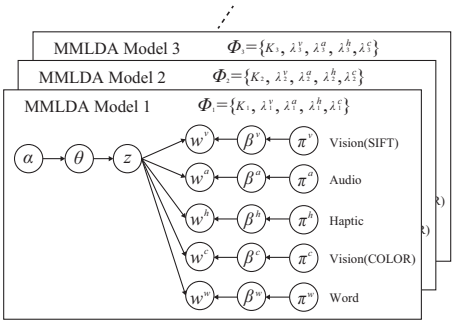


図 3: BoMLDA のグラフィカルモデル

図 4: 実験に使用した 45 物体（枠で囲まれた物体は，単語予測用として用いた）

よって変化する．ここでは, パラメータ ϕ を変化させ, 様々なカテゴリを学習する．すなわち, 複数のモダリティと結びついた概念や, 特定のモダリティと結びついた概念をあらわすモデルが学習される．このように，BoMLDA は様々なカテゴリから形成されている．

2.6 単語が表すカテゴリの選択

BoMLDA では，パラメータ ϕ を変化させたモデルを数多く学習することで，様々なカテゴリを構築した．次に，単語が表すカテゴリとモデルのパラメータの選択を行う．単語とカテゴリの結びつきの強さの尺度として，単語とカテゴリ間の相互情報量を用いる．単語 x^w とモデル ϕ 中のカテゴリ k との相互情報量は以下の式より計算することができる．

$$I(x^w, k|\phi) = \sum_{K\in(k,\bar{k})} \sum_{W\in(x^w, x^{\bar{w}})} P(W,K|\phi) \log \frac{P(W,K|\phi)}{P(W|\phi)P(K|\phi)} \quad (8)$$

ただし，$\bar{k}$ は k 以外のカテゴリを表し，$x^{\bar{w}}$ は x^w 以外の単語を表している．相互情報量とは，二つの確率変数の共有する情報量であり，相互依存の尺度である．したがって単語とカテゴリ間の相互情報量が大きい場合，その単語はそのカテゴリを表現しているといえる．最終的に，単語 x^w が表すモデル ϕ_{x^w} とカテゴリ k_{x^w} は以下の式で選択される．

$$(\phi_{x^w}, k_{x^w}) = \underset{k,\phi}{\operatorname{argmax}}\, I(x^w, k|\phi) \quad (9)$$

2.7 単語の予測

まず，2.6 の手法により選択した単語 x^w と相互情報量の高いモデル ϕ_{x^w} を用いて，未知物体のカテゴリの推定を行う．未知物体のマルチモーダル情報から，学習したパラメータを用いて未知物体がそれぞれのカテゴリに属する確率を計算することになる．未知物体のマルチモーダル情報 w^v_{obs}, w^a_{obs}, w^h_{obs}, w^c_{obs} が与えられた場合，選択されたモデルにおいて，そのカテゴリは $P(z|w^v_{obs}, w^a_{obs}, w^h_{obs}, w^c_{obs}, \phi_{x^w})$ を最大とするカテゴリ z を選択すればよいことになる．従って，未知物体のカテゴリは,

$$\hat{z} = \underset{z}{\operatorname{argmax}}\, P(z|w^v_{obs}, w^a_{obs}, w^h_{obs}, w^c_{obs}, \phi_{x^w})$$
$$= \underset{z}{\operatorname{argmax}} \int P(z|\theta, \phi_{x^w}) P(\theta|w^v_{obs}, w^a_{obs}, w^h_{obs}, w^c_{obs}, \phi_{x^w}) d\theta \quad (10)$$

に よって 決 め る こ と が で き る ． た だ し ，$P(\theta|w^v_{obs}, w^a_{obs}, w^h_{obs}, w^c_{obs}, \phi_{x^w})$ は学習時に推定した β^v, β^a, β^h, β^c, β^w を固定し，前節のパラメータ推定を行うことで求めることができる．

ここで，推定されたカテゴリ $\hat{z}$ が単語と相互情報量の高いカテゴリと一致した場合，つまり，

$$\hat{z} = k_{x^w} \quad (11)$$

図 5: 形成されたカテゴリの例

となる場合，未知物体から単語 x^w が予測されたことになる．最終的に，2.6 の手法により選択された全てのモデルにおいて上記の手法を行い，未知物体から予測される単語を決定する．

3. 実験

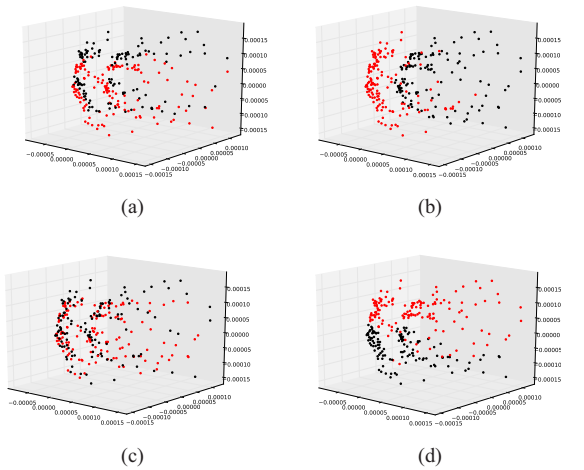
図 1 に示すロボットにより，取得した視覚 (SIFT)・視覚 (色)・聴覚・触覚・単語情報を用いて実験を行った．実験には図 4 に示す 45 個の物体を使用し，カテゴリ分類実験及び単語の予測実験 (学習用物体として，各カテゴリから一つの物体を無作為に抽出した) を行った．なお，単語情報としては図 4 の 45 物体の色や握った感触を表す計 26 種類の単語を用いた．

3.1 カテゴリの学習

まず，各特徴量の重み w^* を 0,300 の 2 段階に変化させ，カテゴリ数は 2〜19 に設定し，BoMLDA の学習を行った．全ての重みが 0 となる場合を除くため，$(2^4 - 1) * 18 = 270$ 個の MLDA から構成されることになる．最終的に，与えられた単語の相互情報量が最大となるカテゴリを選択した結果の一部が図 5 である．物体カテゴリを表すぬいぐるみやゴム人形といったカテゴリが正しく形成できていることがわかる．また，カテゴリ「楽器」には音が鳴る物体が全て含まれており，「楽器」を表すカテゴリが正しく形成できたと言える．さらに，色を表すカテゴリや触覚を表すカテゴリ等，特定のモダリティと結びついたカテゴリも概ね正しく形成できている．

3.2 モデル間の関係の可視化

次に，各モデルの関係を可視化するために，Multidimensional Scaling(MDS) により各 MLDA モデルを 3 次元空間にプロットした．MDS は，多変量解析の一手法であり，各モデル間の距離から，その関係を低次元の空間で表現するものである．しかし，MLDA のモデルでは，各モデル毎にモデル構造



(a) (b) (c) (d)

図 6: MDS によるモデルの 3 次元プロット (各点が一つの MLDA モデルを表し，赤い点が高い重みを示している) (a) 視覚 (SIFT) の重み　(b) 聴覚の重み　(c) 触覚の重み　(d) 視覚 (色) の重み

が異なるため，単純にモデル間の距離を計算することができない．そこで，学習用物体のマルチモーダル情報 x_j^* から単語 x^w が発生する確率を表す確率分布 $P(x^w|x_j^v, x_j^a, x_j^h, x_j^c, \phi)$ の KL 距離をモデル間の距離として用いた．よって，パラメータが ϕ_1 となるモデルと，パラメータが ϕ_2 となるモデル間の距離は，以下のように表現できる．

$$D(\phi_1|\phi_2) = \sum_j \sum_{x^w} P(x^w|x_j^v, x_j^a, x_j^h, x_j^c, \phi_1) \times \log \frac{P(x^w|x_j^v, x_j^a, x_j^h, x_j^c, \phi_1)}{P(x^w|x_j^v, x_j^a, x_j^h, x_j^c, \phi_2)} \quad (12)$$

図 6(a)-(d) が，各モデルを点としてプロットし，視覚 (SIFT)・聴覚・触覚・視覚 (色) の重みが高いものを赤い点として図示したものである．この結果から，この 3 次元空間上において，左側に聴覚の重みが高いモデルが，上方に視覚 (色) の重みが高いモデルが存在していることが分かる．また，視覚 (SIFT) および触覚の重みの高いモデルは大きな偏りは見せず，全体的に散らばった結果となった．これは，色や音に比べてテクスチャ情報や触覚情報は，MLDA における分類に大きな変化を与えないためだと考えられる．例えば，視覚・触覚情報を用いた分類においては，動物の形をした楽器とぬいぐるみは同じカテゴリに分類されるが，聴覚情報を用いた場合は別のカテゴリに分類される．また，色情報を用いた場合は物体概念をある程度無視して色ごとの分類が行われる．

3.3 単語の予測

次に，学習用物体を用いて BoMLDA により学習を行い，予測用物体である未知物体の視覚 (SIFT)・視覚 (色)・聴覚・触覚情報を用いて単語の予測を行った．なお，図 4 の矩形で囲まれた物体が予測用物体である．表 1 に予測された結果の一部を示す．かえるのぬいぐるみから「灰色」やスポンジのボールから「コップ」といった間違った単語がいくつか予測されているものの，概ね正しい単語が予測されていることがわかる．また，10 物体から予測された単語の適合率，再現率及び F 値の平均値はそれぞれ 0.77，0.88，0.81 となった．

表 1: 予測された単語の例

未知物体	予測された単語
	柔らかい，動物，緑，ぬいぐるみ，灰色
	柔らかい，楽器，茶色
	柔らかい，スポンジ，ボール，丸い，コップ，赤
	硬い，楽器，マラカス，赤

4. まとめ

本稿では，ロボットが取得した視覚・聴覚・触覚・単語情報を用い，BoMLDA により多様な概念を形成する手法を提案した．これにより，物体カテゴリだけでなく，色に注目したカテゴリや触覚に注目したカテゴリなど，様々なカテゴリ分類が可能となることを実験を通して明らかにした．また，色カテゴリに注目したモデル，物体カテゴリに注目したモデルなど，様々なモデルを用いての単語の予測を可能とした．さらに，MDS により低次元空間にプロットすることで，それぞれの MLDA の関係の解析を行い，提案手法によるモデル選択が有効であることが示された．

今後さらに実験を進めることで，単純に相互情報量の高いモデルを選択するだけではなく，重視されている特徴量に注視して物体の再学習を行うことにより，形成されるカテゴリの精度向上を行う予定である．また，物体数・カテゴリ数の拡大，人からの教示発話を直接単語情報として用いること，「これ」などの物体の特徴を表現しない機能語の扱い方，学習のオンライン化も今後の重要な課題である．

参考文献

[Rosch 99] Rosch,E.: “Principles of categorization,” Concepts: core readings, pp.189–206, 1999.

[Blei 03] Blei,D.M. *et al.*: “Latent dirichlet allocation,” Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[Araki 12] Araki,T.*et al.*: “Online object categorization using multimodal information autonomously acquired by a mobile robot,” Advanced Robotics, Vol.26, Issue 17, pp.1995–2020, 2012.

[Nakamura 08] 中村ほか: “ロボットによる物体のマルチモーダルカテゴリゼーション,” 電子情報通信学会論文誌 D, vol.91，pp.2507–2518，2008.

[Nagai 12] 長井ほか: “マルチモーダルカテゴリゼーション－経験を通して概念を形成し言葉の意味を理解するロボットの実現に向けて－ション,” 人工知能学会，vol.27，No.6，pp.555–562，2012.

[Nakamura 09] Nakamura,T.*et al.*: “Grounding of word meanings in multimodal concepts using LDA,” in Proc. of IROS, pp.3943–3948, 2009.

[Vedaldi 10] Vedaldi, A. *et al.*: “VLFeat: An open and portable library of computer vision algorithms,” ACM International Conference on Multimedia, pp.1469–1472, 2010.

[Nakamura 10] 中村ほか: “把持動作による物体カテゴリの形成と認識”, 情報処理学会全国大会 2010, 5V-3, 2010