知的情報処理特論

2015/06/04
担当：伊藤
T15M074

5.複数の感覚機能を同時に使った因果関係のモデル化に関する実験

以前の実験では、提案フレームワークは、クロスモーダルメモリの取得および自己組織化マルチモーダル融合表現を利用した安定した行動認識に成功したことを明らかにしました。このセクションでは、他の入力モダリティとして音声信号を組み込むことによって、実験的な設定を拡張します。実験を通じて、提案フレームワークは、環境中の感覚運動経験から複数の感覚機能を同時に使った因果関係を抽出し、そして、取得した因果関係モデルを利用し、感覚の結果を予測する 方法を検討しました。

5.1 提案フレームワークの構築

図11は、提案フレームワークの概略図を示します。三つの独立したディープニューラルネットワーク（すなわち、オートエンコーダ）は音声圧縮、画像圧縮、及び時系列の学習のために利用されます。ロボットの頭部に取り付けられたマイクから取得された音声データは、離散フーリエ変換（DFT）によって前処理されます。音声圧縮ネットワークは（図11の(a)）、取得した音声スペクトルを入力し、中間層から対応する特徴ベクトルを出力します。同様に、画像圧縮ネットワーク（図11(b)）はロボットの頭部に取り付けられたカメラから取得した生のRGBビットマップ画像を入力し、対応する特徴ベクトルを出力します。音声と画像の特徴は、関節角度ベクトルと同期しており、マルチモーダル時間的セグメントが生成されます。これらのマルチモーダル時間的セグメントは、次に時系列の学習ネットワークに渡されます（図11(c)）。したがって、マルチモーダル特徴と再構成されたマルチモーダルセグメントはそれぞれ、中間層と出力層からの出力です。 時系列の学習ネットワークからの出力は、ロボット運動の生成、音声スペクトルの取得、または画像検索のために使用することができます。ネットワークからの関節角度出力は、再スケーリングとモーションを生成するための関節角度指令としてロボットへの再送信がされます。ネットワークはまた、対応する特徴出力を伸張して元の形式で取得した音声スペクトルや画像を再構成することができる。なぜなら音声圧縮ネットワークと画像圧縮ネットワークは、中央隠れ層における特徴ベクトルを介して、入力から出力への特徴マップを作るためです。

5.2実験準備

提案機構のクロスモーダルメモリ検索のパフォーマンスは、最初の実験で使用したのと同じロボットでベルリンギングタスクを行うことにより評価され、次のようにセットアップされます。：表面の色や音のピッチのいずれかによって識別することができる3つの異なる机上のベルは、実験のために準備されます。色およびピッチ表記との対応は、図12(a)に示されています。各ベルリンギング試験のために、2つのベルが選択され、ロボットの前側に並んで置かれました。そして、 2つのベルのどちらか一方がベルの上部にあるボタンを押すことで鳴らされます。手の限られたリーチのために、ベルの各側に対応する腕でのみ鳴らさせることができます。図12(b)に示すように、6つの可能なベルの配置の組み合わせがあります。タスク構成の下で、少なくとも二つの異なるモダリティからの情報が右ベルリンギングの状況を判断するために必要なことに注意してください。実際には、RGB画像・音声情報・関節角度情報それぞれ単独では、予測や決定ができない。

6つの異なるベルの配置構成の下で左右のベルを打つ動きを生成することにより、12種類のマルチモーダル時系列データセットを記録します。ベルを打つ動きに応じた腕関節角度シーケンスは最初の角度補間とターゲットの姿勢によって、データセットが生成されます。パルス符号変調（PCM）音声データ(16 kHzのサンプリングレート、16ビットの深さ)は、ロボットの額に取り付けられたマイクを備えた単一のチャネルで記録されます。画像フレームと両腕の関節角度は、複製された画像フレームを含む約66Hzで記録されています。画像と関節角データと一緒に音声データを同期するには、242サンプルのハミング窓や重複なしでウィンドウシフトの242サンプルでDFTによって前処理されます。320 × 200ピクセルの一部の領域は、元の320 ×240の画像からトリミングされ、私たちのコンピュータ環境のメモリリソースの限度に調整するために、40 × 25ピクセルにサイズ変更されます。関節角度データ入力のために、（肩から手首までの）腕の自由度10種が使用されます。モーションシーケンスの結果の長さは約3秒ごとに相当する約200ステップでした。マルチモーダル時系列学習のために、一つの入力として元の時系列から30ステップの連続セグメントを使用します。ワンステップ時間窓をスライドすることで、連続したデータセグメントが生成されます。 表3は、データセットおよび関連する実験パラメータをまとめたものです。音声特徴および映像特徴学習の両方について、同じ12層のディープニューラルネットワークが使用されます。時系列の学習については、10層のネットワークが使用されています。各ケースにおいて、デコーダアーキテクチャは、対称のオートエンコーダをもたらすエンコーダの鏡像です。ネットワーク構造のパラメータは、経験的に、[27]及び[33]などの以前の研究に基づいて決定されます。次のように2つのネットワークの入力と出力の寸法が定義されています：入力には、(1)単一のベクターにサウンドスペクトル（すなわち、 242次元）の連続

した4段階のシーケンスを結合することによって定義されている音声特徴学習のために968次元、(2)RGBの色毎の画素の40 × 25の行列により定義される画像特徴学習のために3000次元。そして(1)から30次元の音声特徴ベクトルと(2)から30次元の画像特徴ベクトル、10種の関節角度からなる70次元のマルチモーダルベクトルの30段階のセグメントによって定義される時系列学習のために2100次元。特に、時系列の学習ネットワークの中心隠れ層の、ノードのいくつかの数字(30,50,70,100)を比較しました。音及び関節角度入力から画像検索の性能を評価することによって、100ノードが所望のメモリ再構成精度を達成するために必要であると結論しました。

5.3 音と動きのシーケンスからの画像シーケンスの取得

音と関節角度入力シーケンスから画像シーケンスを生成することによって、クロスモーダルメモリ検索性能の評価実験を行いました。なお以下の結果が、シーケンスのステップ数は、記録されたデータステップではなく生成ステップを示しています。具体的には、シーケンスの最初の段階を取得するために、生成工程の開始前の 29 段階のデータが使用されます。図 13 は、音及び関節角度入力から画像生成結果の一例を示しています。配置されたベルの色は任意の音声入力を取得する前の導出ではないので、ステップ 1 では、検索された画像中のベルが任意に着色されています。関節角度の入力データは右アームがベルを打つために使用されようとしていることを示すため、ロボットの右手の画像が既に取得した画像に含まれています。ステップ 31 と 61 で、ベルが鳴らされ、対応する音声スペクトルを取得します。そして、タスクの構成が明らかになり、右側の鳴ったベルがピッチ「F」を有するという情報は、緑色と相関しています。このように、検索された画像の右のベルの色は、音と関節角度の情報を関連付けることによって、ランダムな初期化された一つから緑色にします。逆に、左側に配置されたベルの認識のための音声入力から情報が何も取得されないため、検索された画像の左側のベルの色は、実行時に安定していませんそれにもかかわらず、検索された画像は、鳴ったベルの色（すなわち、緑）が識別された場合、他のベルの色が残りの二つの色（すなわち、赤または青）から選択されることを示しています。この結果は、2 つのベルの色は常に異なっている現在のタスクの設計を反映しています。ステップ 91 付近から、ベルの音が減衰し始めると、姿勢の初期化によるマニピュレータの作動ノイズが優位なります。

5.4画像検索性能の定量的評価

提案モデルが、画像、音声、運動モダリティ間の因果関係をモデル化することに成功し

たかどうかを調べるために定量的に評価実験を行いました。ネットワークのための10種類の初期モデルのパラメータ設定を作成し、ベルの配置とベルを打つ運動パターンの12種の組み合わせからなる同一のデータセットを学習する実験を複製しました。10種の学習結果のクロスモーダル画像検索の結果、画像シーケンスの120パターンが得られました。画像検索のパフォーマンスは元の対応する領域に対して、検索された画像に手動で左右のベル領域を設定し、平均二乗(RMS)誤差によって定量化されます。図14は、音声のパワースペクトルの最大値と関節角度配列に対応して表示された画像検索誤差の時間変化を示しています。画像検索エラーの評価結果として、左側のベルが鳴らされるときは、右側のベルよりエラーが小さくなります。エラー軌跡の時間変化はベルの音を取得した後、検索エラーが減少することを示しています。ベルが鳴らされているときにエラー軌跡の形状は、二つのグラフは対称ではありません。左ベルが鳴らされるとエラーがあっても腕の姿勢の初期化後にその値を維持します。右のベルが鳴らされたとき、エラーは腕の姿勢の初期化後に増加します。これはアクチュエータの違いに起因し、右腕は左腕よりも騒音を生成します。

5.5 生成した動きと検索されたベル画像間の相関

左右の領域の画像検索性能の差の有意性を評価するために、シーケンスのステップ60で、画像検索エラーのt検定を行っています。その時間ステップで、腕がさがり、手はベルを触ります。したがって、画像検索に作動ノイズの影響はありません。鳴らしている方の絵は本物に近い元画像と差が小さく、t検定より統計的に誤差が小さいと言える。結果は、ベルの色と音の間の関連があるように、画像中のベル領域と物理的な運動との間の空間的な相関関係が正しくモデル化されていることを示しています。したがって、画像、音、動きの間に取得されたモダリティの因果関係モデルは、画像検索のために利用されます。

5.6 マルチモーダル特徴空間の可視化

最後に、時系列の学習ネットワークによって取得されたマルチモーダル特徴空間の分析を行いました。10種の複製された学習結果の中で、ベルリンギングシーケンスの12パターンが入力されると、ネットワークの中間層で同一の活性化パターンを記録しました。取得した主成分で定義された三次元空間に得られた100次元の特徴ベクトル系列を投影する主成分分析（PCA）を適用しました。図16（a）は、ロボットの運動パターンは、第1主成分および第2主成分からなる二次元空間で表現されることを示しています。この分析の結果は、複数のモダリティ間の因果関係が学習できたことを示しています。

補足

T 検定
帰無仮説が正しいと仮定した場合に、統計量が t 分布に従うことを利用する統計学的検定法の総称である。 2 つの平均値の差の統計的有意性を検討する。

帰無仮説
証明したい仮説の反対の仮説。

t 分布
連続確率分布の一つであり、正規分布する母集団の平均と分散が未知で標本サイズが小さい場合に平均を推定する問題に利用される分布。