Determining Correspondences of the Audio-Visual Events Caused by Multiple Movements

（複数運動で生じた視聴覚事象の対応付け）

人間は視覚，聴覚，触覚などの多種類の感覚系を持つ．そして，複数の異種感覚器で得られた情報を統合することによって，周囲環境についてより詳しい情報を獲得する．異種感覚情報の統合には，異種感覚器で観察した事象の対応付けが必要となる．生後１８～２０週間の新生児は，発話の音響的特徴（スペクトル）と口の動きによる視覚的特徴（網膜への光刺激の動き）との間で対応付けを行う能力を有することが確認されている．同時に，新生児が口の動きの映像に対して音声を真似る傾向も報告されている．また，人間は複数の人が歩いているシーンで，耳から聞こえてくる足音が，誰の足の動きによって生じているのかを視覚的に特定することができる．このように人間は意識することなく異種感覚情報を統合して状況分析を行う．

センサフュージョンの目標の一つは，人間のこのようなメカニズムを工学的に実現することである．すなわち，複数の異なる感覚情報を統合する機能を持つ知的システムを実現することである．本論文では，人間の代表的な感覚機能であり，重要な役割を占める視覚と聴覚に着目し，それらから得られる情報の統合，すなわち，物体の運動と音の対応付けを考える．

高騒音の環境で音声認識する場合，騒音の影響のため聴覚情報が不十分な場合は，視覚からの手掛かりによって聴覚情報を補うことができる．視聴覚事象の対応付けは視聴覚障害者のための感覚代行・感覚補助システムの実現，複数異種感覚器を持つ知能ロボットセンシングシステムなどに応用できる．

視聴覚情報の統合を用いて，計算機による環境の理解と知識の学習に関する研究はいくつかある．しかし，それらの研究は一つの音と一つの動きあるいは複数の動きしか対応できなく，音源位置が変化する場合や運動を観察する視点が変化する場合には対応できなく．かつ，オフラインで処理である．本論文は，物体に固有な知識を用いず，一般的な物理法則によって，複数運動で生じた視聴覚事象を対応付けることを目的とする．すなわち，複数の動きと複数の音とを対応付ける．視野内で音源位置が変化しない場合（固定音源），視野内で音源の位置が変化する場合（移動音源）と複数音源が同時にカメラの視野内に存在しない場合について，複数の視聴覚事象の対応付けが可能となる３つの手法を提案する．以下にその手法について述べる．

第一の方法は，固定音源において，物体固有の知識を利用せず，一台のカメラと一台のマイクで観察した複数の運動の視聴覚事象を対応付ける．この手法は，対象が繰り返し運動と非繰り返し運動の場合のいずれでも適応できる．対応付けの手掛かりとして，音の発生と運

動変化が同時に生じること，音の繰り返しと運動の繰り返しが似ていることを利用する．音響信号のオンセットでの周波数成分の類似性を用いて，音響信号をグルーピングし，各音源のオンセットの時系列を求める．音のオンセットの時刻に対応する画像間の差分から，運動物体が存在する領域を抽出する．繰り返し運動の場合，領域のエッジの画素数の時系列と音のオンセットの時系列の相関を計算し，音のオンセットでの周波数成分をそれと高い相関を有する運動の領域に対応付ける．非繰り返し運動の場合，音のオンセットでの周波数成分を運動方向が変化した運動ベクトルに対応付ける．

第二の方法は，複数の運動物体が存在して，物体が全体移動すると音源の位置が変化する場合において，一台のカメラと一台のマイクで観察した複数の視聴覚事象を対応付ける．対応付けの手掛かりとしては，音の発生と運動変化が同時に生じること，音の発生と運動パターンの開始とが同じであること，あるいは音の繰り返しパターンと運動の繰り返しパターンとの類似性を利用する．音源の位置が変化することによって，観察した運動の位置は時刻ごとに異なる．しかし，運動の時空間不変量は同じ運動であれば，位置が異なっても変化しない．そこで，時空間不変量を利用することにより，その運動を識別できる．同一の運動であると判断されたものをグルーピングし，視覚時系列を求める．また，第一の方法と同じように聴覚時系列を求める．これらの相関値を算出することによって，音のオンセットでの周波数成分と運動の時空間にマッピングされた特徴点から算出される量（時空間不変量）の時系列パターンとを対応付ける．

第三の方法は，複数の運動物体が同時にカメラ視野に存在しない場合において，一台のパン制御可能なカメラと 2 台のマイクで観察した視聴覚事象を対応付ける．対応付けの手掛かりとして，音の発生した位置と運動物体の位置が同じであること，音の発生と運動物体の変化の時刻が同じであること，音の繰り返しパターンと運動の繰り返しパターンが類似していることを利用する．運動物体を観察するため，2 台のマイクで計測した音響信号から音源の方位を計算し，カメラを制御する．音源定位の誤差とカメラ制御の誤差により，運動物体が常にシーンの中央に計測されることはない．また，音源の位置はシーン内で変化する．そこで時空間不変量を用いて運動を識別する．視聴覚時系列に対して相関を算出することによって，音のオンセットでの周波数成分と運動の時空間不変量の時系列に対応付ける．

以上，本論文では，複数の運動で生じた視聴覚事象を一般的な物理法則によって，対応付ける方法を提案し，その有効性を実験により示した．今後の課題は，自律的に学習した運動の視聴覚事象を蓄積し，物体認識及び環境理解を可能とするシステムを作成することである．