実世界センシングの最前線
―イベントカメラと音響信号によるシーン・人物計測―

慶應義塾大学　理工学部 電気情報工学科
青木義満

1．はじめに

カメラから得られる画像・映像から意味のある情報を頑健に抽出・計測・認識する「画像センシング」技術は，深層学習の登場以降，目覚ましい発展を遂げてきた。物体検出やセグメンテーション，人物の姿勢推定といった基本タスクの精度は実用水準に達し，近年では大規模データで事前学習された基盤モデルの登場により，コンピュータビジョン（CV）研究は新たな段階に入りつつある。

しかし，こうした発展の大部分は「通常の RGB カメラで撮影されたフレーム画像」を入力とする枠組みの中でのものである。実世界のセンシング課題に向き合うと，この枠組み自体の限界に突き当たる場面が少なくない。高速に運動する対象はモーションブラーに埋もれ，逆光や夜間のシーンではダイナミックレンジが足りない。そもそもカメラで人を撮影すること自体が，プライバシーの観点から受け入れられない現場も多い。認識アルゴリズムをいくら高度化しても，センサに届く情報そのものが失われていれば打つ手はない。実世界センシングの最前線では，「何を推定するか」だけでなく「何で，どのように観るか」―すなわちセンシングのモダリティそのものが問い直されているのである。

筆者らの研究室では，画像センシング技術により実世界の人物・物体・シーンの状態を計測・認識・理解することを目指して研究を進める中で，この問いに正面から取り組んできた。本稿では，その最近の成果から二つの方向性を紹介する。一つは，通常の RGB カメラを超える視覚センシングとしてのイベントカメラの研究である。コントラスト最大化と呼ばれる基本原理に基づく一連の研究展開と，その最新成果であるイベントベース 3 次元復元手法 GPERT（CVPR 2026 Highlight 採択）[1]を解説する。もう一つは，視覚情報を用いないセンシングへの挑戦であり，環境音楽を信号源として人の 3 次元姿勢を推定する BGM2Pose[2]を紹介する。イベントカメラが「時間」の次元でカメラの限界を超える試みだとすれば，音響センシングは「モダリティ」の次元で視覚の限界を超える試みである。いずれも，センサ・信号の物理特性まで遡って機械学習と一体で設計するアプローチであり，光学・光技術に携わる本誌読者の皆様にも興味を持っていただける内容と考えている。

2．イベントカメラによる実世界センシング

2-1　イベントカメラの原理と特長

イベントカメラ（イベントベースビジョンセンサ）は，生物の網膜の仕組みに着想を得た新しい視覚センサである[3]。従来のフレームカメラが一定周期で全画素の輝度値を同期的に読み出すのに対し，イベントカメラは各画素が独立に対数輝度の変化を監視し，変化量が閾値を超えた瞬間に，その画素位置・時刻・変化の極性（明→暗／暗→明）の組を「イベント」として非同期に出力する(図 1)。出力は（画素座標，時刻，極性）の 3 つ組の系列であり，時刻はマイクロ秒精度で記録される。

この撮像原理は，フレームカメラと比較して際立った特長をもたらす。第一に，マイクロ秒オーダーの時間分解能である。イベントは画素ごとに変化検出の瞬間に出力されるため，フレームレートの概念に縛られず，高速な運動をモーションブラーなしに捉えられる。第二に，120 dB を超える高ダイナ