ドライブレコーダデータに対するヒヤリハット発生対象分類

山本 修平[1]　倉島 健[1]　戸田 浩之[1]

概要：ドライブレコーダで記録された映像やセンサデータは，交通事故時の証跡に使われるだけでなく，交通違反や交通事故等の危険な状況に直面した際のいわゆる「ヒヤリハット」データとして，ドライバーの安全運転教育に利用されている．大量のドライブレコーダデータの中から，安全運転教育に資するヒヤリハットデータの利活用のため，各データに人間が理解可能なラベルを付与し，これらのデータを簡便に検索できるようになることが期待されている．本論文では，ドライブレコーダデータに対して，非ヒヤリハットを含めたヒヤリハットの発生対象を分類する手法を提案する．提案手法は，映像とセンサ系列の時間的遷移を考慮した特徴変換と，前方映像からの物体検出結果をグリッド空間へ埋め込んだ行列に対する特徴変換によって 2 種類の特徴ベクトルを得る．この特徴ベクトルを組み合わせ，マルチタスク学習によって，より簡単なサブタスクの推定結果を活用してメインタスクであるヒヤリハット発生対象を推定する．実際のドライブレコーダデータを用いた評価実験の結果，物体検出結果のグリッド空間の埋め込みとマルチタスク学習を組み合わせる方法が，ベースライン手法に比べて高い分類性能を示すことを明らかにした．

Traffic Near-miss Target Classification on Event Recorder Data

SHUHEI YAMAMOTO[1]　TAKESHI KURASHIMA[1]　HIROYUKI TODA[1]

1. はじめに

ドライブレコーダとは，車内に設置され，前方映像，速度，加速度，ウィンカ操作，ブレーキ操作などの車両運行状況を記録する装置である [15]．全時刻の情報を記録することは，記憶容量などの問題もあるため，実際には加速度に一定の閾値を設け，車両に何らかの衝撃が見られたと考えられるタイミングを検知し，その前後十数秒を記録することが多い．このように記録されたデータを，本論文ではイベントデータと呼ぶ．このようなデータは，交通事故の証拠としてだけでなく，交通違反や交通事故等の危険な状況に直面した際の，いわゆる「ヒヤリハット」に関するベントデータを，ドライバーに実際に視聴してもらったり [32]，視聴しながら危険予知訓練をするなど [5]，安全運転を促進することにも活用されている．また，法人営業車両などを対象に，走行中の運行車両からヒヤリハットシーンをリアルタイムに検出し，運行管理者に通知することによって，頻繁に危険運転をするドライバーの事故を未然に防ぐことにも利用されている [35]．実際のヒヤリハットシーンとなった前方映像の画像を**図 1** に示す．GfK ジャパン社の調査によると，ドライブレコーダの国内販売台数は 2017 年に前年比 38%増の 109 万台となっており，今後もさらなるヒヤリハット映像の収集が見込まれる [37]．

図 1　ドライブレコーダで記録された実際のヒヤリハット例．左図では，交差点に侵入する車に対して，右図では，横断歩道をわたる自転車に対して急ブレーキし，ヒヤリハットを起こしている．

一方，ドライブレコーダで記録された全てのイベントデータがヒヤリハットに関するものではない．東京農工大学スマートモビリティ研究拠点の調査によると，たとえば，道路の激しい起伏が原因で車両に大きな衝撃が加えられ記録されたものや，狭い道路での方向転換ために行う小刻みな加速が原因で記録されたもの等，非ヒヤリハットに該当するイベントデータも全体の 70%近く含まれている [30]．

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所

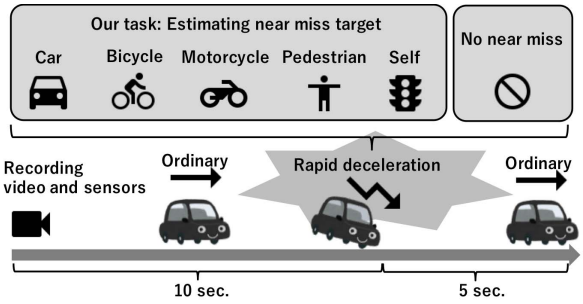

図 2　本論文で取り組むタスクの概要

また，実際の安全運転教育では，ヒヤリハットの発生対象（例えば，バイクや歩行者とのヒヤリハットなど）毎のように，イベントデータを類似イベント毎にまとめることも望まれている．しかしながら，収集された膨大なイベントデータの中から，ヒヤリハットに該当するイベントデータを人手で抽出し，更にその発生対象等をラベリングすることは，その判定者に多くの労力や注意力を要する．このため，これらの蓄積されたイベントデータの中から，自動的にヒヤリハットを含むイベントデータを抽出しその種類に応じてラベリングすることは，安全運転教育にかかるコストを低減し，様々なケースを教育に活用できることから，それを促進する効果が期待される．そこで本論文では，ヒヤリハットが起こった際の発生対象に注目し，イベントデータの集合に対して，ヒヤリハットの有無，ヒヤリハットがある場合には適切なヒヤリハット発生対象のラベルを付与するタスクに取り組む（**図 2**）．

これまでも，ヒヤリハットを自動的に検出することに取り組んでいる研究がいくつか存在する．ドライブレコーダに含まれるセンサ信号に対して，経験的にルールを設定しヒヤリハットの抽出を試みている研究 [31] や，センサデータを適切な特徴量に変換して右折時に限定したヒヤリハット発生形態を分析し，その分析結果に基いてヒヤリハット検出まで行う研究 [34] などがある．しかしながら，これらはセンサデータに基づくヒヤリハット検出を目的としており，その発生対象のラベリングまでは行っていない．特定のヒヤリハット発生対象に限定した研究では，前方映像に対して人物検出を活用して歩行者とのヒヤリハットを検出する研究 [11] や，深層学習を活用して歩行者とのヒヤリハットを危険度と共に推定する研究 [22] があるが，これらは歩行者以外の発生対象を扱っていない．

著者らは，これまでにドライブレコーダに含まれるセンサと映像からなるマルチモーダルデータを用いた，教師あり深層学習に基づくヒヤリハット検出手法を提案している [33]．ここでは，センサと映像をそれぞれ Fully-Connected Neural Network（FC）と Convolutional Neural Network（CNN）[6] で特徴変換し，Recurrent Neural Network（RNN）[17] で時間的な遷移をモデル化している．実際のイベントデータを用いた評価実験の結果，提案手法が高い精度でヒヤリハット検出（ヒヤリハットの有無の判定タスク）できることを明らかにしている．従来手法はヒヤリハット発生対象を推定するための多クラス分類に容易に拡張可能であるが，素朴な拡張では精度が低下することが予想される．その理由として，以下の 2 項目がある．

課題 1　ヒヤリハット検出タスクでは，前方映像内に対象物が存在することがわかればある程度の判定ができたため，画像を CNN で解析することで得られた特徴量を用いれば十分であった．しかし，対象物のラベリングタスクにおいては，何の物体に対するヒヤリハットであるかを判別する必要があり，単純な CNN を用いるだけでは，その判別に十分な特徴量を得られず，精度が低下すると考えられる．このため，前方映像に映る対象物を判別するための特徴量を別途抽出する必要がある．

課題 2　本論文で扱うヒヤリハット発生対象分類では，各イベントデータに対してヒヤリハットの有無の判定に加え，ヒヤリハットと判定されるデータに対してその発生対象を推定する．このように，推定対象とするラベル集合に階層構造が含まれる場合に，単純な多クラス分類モデルではその情報が活用されず，精度が低下すると考えられる．このため，階層構造を考慮し，各イベントデータにヒヤリハットの発生対象を推定する必要がある．

このような課題を解決するため，本論文では，物体検出結果とマルチタスク学習を活用したヒヤリハット発生対象の，教師あり深層学習に基づく分類手法を提案する．提案手法は大きく 3 つの要素から構成される．1 つ目は画像と物体検出結果，センサからなる時系列データを RNN 等で適切な特徴ベクトルに変換する，Temporal Encoding Layer である．2 つ目は前方映像中に映る対象物を物体検出を用いて抽出し，その種類や映像中の出現位置，また自車との位置関係を考慮して，前方映像の領域をグリッドで分割した空間に埋め込み，深層学習によって特徴変換する Grid Embedding Layer である．3 つ目はこれら 2 つの特徴変換から得られた特徴ベクトルを組み合わせ，ラベルの階層関係を活用して抽出した 2 つのサブタスクを推定した後に，その推定結果を用いてヒヤリハット発生対象タスクを推定する Multi-task Layer である．それら 3 つのタスクにおける誤差を組み合わせた目的関数を最小化するよう，ニューラルネットワークを最適化する．

本論文の貢献は以下の 3 項目である．

(1) ドライブレコーダのように Egocentric な視点から撮影された画像において，物体検出によって得られた対象物の出現位置や自車との距離関係を考慮して，特徴変換する手法を提案した（課題 1 へのアプローチ）．

(2) 多クラス分類タスクにおいて，クラス間に階層関係が仮定できる場合に，より簡単なサブタスクを抽出し，

そのサブタスクの推定結果をメインタスクに活用する，マルチタスク学習の枠組みで分類モデルを提案した（課題 2 へのアプローチ）．

(3) 上記 2 つの提案要素に関して，それぞれの有効性と，それらを組み合わせたときの有効性を，ヒヤリハット発生対象分類タスクを通じて明らかにした．

以下，本論文の構成を示す．**2 章**ではヒヤリハットの発生対象分類の関連研究について説明する．**3 章**では本論文で扱うイベントデータの構造やその前処理などを述べる．**4 章**では深層学習を活用したヒヤリハット発生対象分類の手法について説明する．**5 章**では実際のドライブレコーダで記録されたイベントデータを用いた評価実験によって提案手法の有効性を評価し，**6 章**では本論文のまとめと今後の課題について述べる．

2. 関連研究

本章では，ドライブレコーダからのヒヤリハット検出を扱う関連研究と，ドライブレコーダを用いたその他の関連研究について述べ，本研究の位置付けを整理する．

ドライブレコーダからのヒヤリハット検出を対象とする研究として，速度と加速度に経験的にルールを設定し，閾値調整によってヒヤリハットを検出する研究 [31] や，センサデータを適切な特徴量に変換して右折時に限定したヒヤリハット発生形態を分析し，右折時のヒヤリハット検出に取り組む研究 [34] がある．また，前方映像データを扱ったヒヤリハット検出としては，前方映像から歩行者検出を行い，時車位置と映像中の検出領域との距離を計算して，歩行者とのヒヤリハットを検出する研究 [11] や，深層学習を用いて歩行者検出をサブタスクで推定しつつ，ヒヤリハットを 3 段階の危険度と共に検出する研究 [22] がある．これらの研究は，センサデータや映像データを用いてヒヤリハット検出をしているが，本研究が目指すヒヤリハットの発生対象分類は扱っていないことから，本研究とは異なる．

ヒヤリハット検出以外にも，ドライブレコーダは様々なタスクで用いられている．センサデータに含まれるドライバの運転操作を用いたものでは，ヒヤリハットの起こりやすい潜在リスクを持つ道路の推定をする研究 [14], [36] や，ヒヤリハットを起こしやすい危険運転ドライバの分類に関する研究 [27], [28] がある．車両運行中の前方映像データは，深層学習技術の発展により，自動車の自動運転に関する研究で幅広く利用されている．自動運転タスクでは，前方映像に基いてハンドルの操舵角を予測する研究 [1], [12]，車の直進，左折，右折，車線変更，停止などの運転操作を予測する研究 [24]，同様のタスクに車内映像も用いた研究 [8], [9]，交通事故を回避するため，前方で発生する交通事故を予測する研究 [3] などがある．これらはいずれも深層学習に基づく手法を提案しており，画像解析は CNN，時系列モデリングは RNN において有効性の知られている，Long Short-Term Memory（LSTM）[7] を用いている．本論文でも CNN や LSTM を用いたモデル化に取り組むものの，上記のタスクでは各時刻での運転操作やハンドルの操舵角，交通事故の起こりやすさを予測することに対して，本研究では入力の最終時刻でヒヤリハットの発生対象を分類する点で，これらの研究とタスクや提案するモデルが異なる．

表 1 本論文で用いる主な記号

記号	説明
C	非ヒヤリを含む，ヒヤリハット発生対象のラベル数
t_a	非ヒヤリを含む，ヒヤリハットの発生対象を識別する正解ラベル
t_b	ヒヤリハットと非ヒヤリを識別する正解ラベル
t_c	非ヒヤリを除く，ヒヤリハットの発生対象を識別する正解ラベル
T	イベントデータのフレーム数
V	物体の種類数
N^t	t フレーム目の画像から検出された総物体数
$\mathbf{o}_n^t, \mathbf{b}_n^t, p_n^t$	t フレーム目の画像から n 番目に検出された物体の 1-of-K 表現ベクトル，領域ベクトル，検出確率
$\mathbf{e}^t$	t フレーム目の画像の物体検出スコアのベクトル
H, W	オリジナル画像の縦幅，横幅
G_h, G_w	物体検出結果を埋め込むグリッド空間の縦，横のグリッド数
$\mathbf{G}$	物体検出結果を埋め込むグリッド特徴行列
L_a, L_b, L_c	正解ラベル t_a, t_b, t_c に対する推定結果の誤差
L	L_a, L_b, L_c の結果を組み合わせた目的関数
β	マルチタスク学習におけるサブタスクの重みを調整するハイパーパラメータ
U	Fully-connected layer のユニット数
$\mathbf{W}_a, \mathbf{b}_a, \mathbf{u}_a$	Temporal Encoding Layer における Soft Attention のモデルパラメータ
$\mathbf{W}_g, \mathbf{b}_g, \mathbf{u}_g$	Grid Embedding Layer における Soft Attention のモデルパラメータ

3. 準備

本章では，ヒヤリハット発生対象分類をするための，ドライブレコーダで記録されたイベントデータの前処理を述べる．また，本論文で用いる主な記号を**表 1** にまとめる．

3.1 データの概要

ドライブレコーダは，加速度に一定の閾値を設け，その閾値を超えたときを基点に前後十数秒を記録する装置である[*1]．このように記録されるイベントデータは，総フレーム数 T からなるセンサ集合と画像の系列から構成され，時間的な同期がとられているものとする．センサ集合は，数次元（加速度やスピードなど）のベクトルであり，これらの次元間の値のスケールを統一に扱うため，入力された系

[*1] 一般に販売されているドライブレコーダでは，加速度閾値を 0.5G，閾値を超えた時刻を中心に前 10 秒，後 5 秒を記録するものが多い．

列データに対してそれぞれの次元で平均 0，標準偏差 1 になるように正規化する．

3.2 物体検出の適用

ヒヤリハットの発生対象をより高精度に推定するため，本論文では前方映像の画像 I^t $(t = 1, 2, \cdots, T)$ に対し物体検出を行い，その結果を利用する．物体検出処理は高い認識精度を示すことで知られている，事前学習された YOLO9000 [19] を使用する．画像 I^t から得た物体検出結果は，N^t 個の検出されたオブジェクトを持ち，各オブジェクトの ID を総物体種類数 V 次元の 1-of-K 表現で持つベクトル $\mathbf{o}_n^t$，各オブジェクトの領域情報を持つベクトル $\mathbf{b}_n^t$，及び各オブジェクトの検出確率 $p_n^t \in (0, 1]$ からなる $(n = 1, 2, \cdots, N_t)$．ここで，領域情報ベクトル $\mathbf{b}_n^t$ は，その左端，上端，右端，下端の 4 つの画像中の座標を持つ $(\mathbf{b}_n^t = \{b_n^t.left,\ b_n^t.top,\ b_n^t.right,\ b_n^t.bottom\})$．また，1 つの画像から同じ物体名を持つオブジェクトが複数得られることもある．

3.3 アノテーションラベルの再整理

教師あり学習で利用するイベントデータには，ヒヤリハットの発生対象を識別する正解ラベル $t_a \in \{1, 2, \cdots, C\}$ のいずれかがアノテーションされている．特に本論文では，$t_a = 1$ であるときを非ヒヤリハットとして扱う．C は非ヒヤリハットを含むヒヤリハットの発生対象のラベル数である．**4 章**で提案するマルチタスク学習のため，ヒヤリハット発生対象のアノテーションラベル t_a の再整理によって，2 種類の追加正解ラベルを得る．1 つ目の追加正解ラベルは，ヒヤリハットか否かを判別するためのラベル $t_b \in \{0, 1\}$ であり，これは $t_a = 1$ であるときに $t_b = 0$，それ以外のときに $t_b = 1$ として得られる．2 つ目の追加正解ラベルは，ヒヤリハット発生対象のいずれかを判別するためのラベル $t_c \in \{1, 2, \cdots, C-1\}$ であり，これは $t_c \leftarrow t_a - 1$ によって得られる．$t_c = 0$ の場合は非ヒヤリハットであるので，ここの追加正解ラベルでは欠損値として扱う．

4. 提案手法

4.1 概要

本章では，本論文で提案する深層学習を活用したヒヤリハット発生対象の分類手法について述べる．提案手法は大きく 3 つの構成要素からなる（**図 3**）．1 つ目は，総フレーム T からなるセンサ，画像，物体検出結果を各時刻で特徴変換し，時間的な遷移を RNN でモデリングする Temporal Encoding Layer である．2 つ目は，前方映像に対する物体検出結果について各物体の検出領域を考慮し，前方映像の領域をグリッドで分割した空間（グリッド空間）に特徴埋め込みをし深層学習により特徴変換する Grid Embedding Layer である．3 つ目は，Temporal Encoding Layer と Grid Embedding Layer から得られた特徴ベクトルから，サブタスクの推定結果を活用しヒヤリハット発生対象を推定する Multi-task Layer である．以下，それぞれを **4.2 節**，**4.3 節**，**4.4 節**で順に説明する．

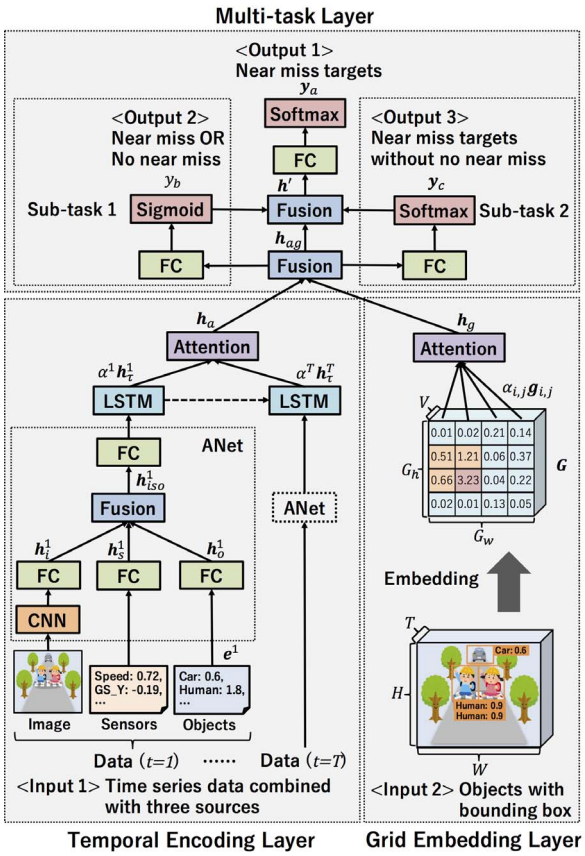

図 3 提案手法の概要図

4.2 Temporal Encoding Layer

ここでは，入力の時系列データ $t = 1, 2, \cdots, T$ に関して，画像，センサ，物体検出結果に分け特徴変換した後，RNN によって時間的な遷移を考慮した特徴ベクトルを得る．具体的には，以下のように処理を行う．

画像データ処理ユニット．前方映像中の天候や道路状況など大まかな特徴を得るために，前方映像の各画像データをニューラルネットワークによってエンコードする．ここでは，ImageNet [20] と Places365 [29] のデータセットで Pre-training した 2 種類の GoogLeNet [23] を用意し，それぞれ 1,000 次元，365 次元の画像特徴ベクトルを抽出する[*2]．FC を通じて U 次元の特徴ベクトルにさらにエンコードする．t フレーム目から得られた特徴ベクトルを $\mathbf{h}_i^t$ と表記する．

センサデータ処理ユニット．車両の運行状況に関する特徴を得るために，平均 0，標準偏差 1 に正規化されたセン

[*2] ImageNet は画像中の物体認識，Places365 は画像中の場所認識のためのデータセットである．GoogLeNet は，このような画像認識で有効性の知られている DNN 構造である．

サデータを，FC を通じて画像データと同様に U 次元の特徴ベクトルにエンコードする．t フレーム目から得られた特徴ベクトルを $\mathbf{h}_s^t$ と表記する．

物体検出データ処理ユニット．前方映像中の障害物や道路標識の有無など細かな特徴を得るために，物体検出結果を簡単なベクトル表現に変換して利用する．ここでは，画像中にどのような物体がどれくらい存在するかを考慮するため，物体検出結果における物体名を識別するための ID ベクトル $\mathbf{o}_n^t$ と，その検出確率 p_n^t を用いて，総物体種類数 V からなるベクトル $\mathbf{e}^t = \{e_1^t, e_2^t, \cdots, e_V^t\}$ を新たに生成し入力する．オブジェクト j の特徴量は，

$$e_j^t = \sum_{n=1}^{N_t} o_{n,j}^t \cdot p_n^t, \tag{1}$$

として得る．すなわち，画像中に同じ物体が複数検出された場合は，その物体の出現度合いを高めるために，その物体の検出確率 p_n^t を合計してベクトルを生成する．$\mathbf{e}^t$ を FC を通じて U 次元の特徴ベクトルにエンコードする．t フレーム目から得られた特徴ベクトルを $\mathbf{h}_o^t$ と表記する．

時系列モデリングユニット．画像，センサ，物体検出結果を適切にエンコードした特徴ベクトルを結合し ($\mathbf{h}_{iso}^t = [\mathbf{h}_i^t; \mathbf{h}_s^t; \mathbf{h}_o^t]$)，FC を通じて $3U$ 次元から U 次元の特徴ベクトルに変換した後，LSTM ユニットを持つ RNN へ入力する．LSTM は系列データの長期的な依存関係を学習できることから，ヒヤリハットの発生対象分類に必要な時間的な状態遷移のモデル化を期待できる [7]．さらに，Soft Attention によって，より直接的に全時刻の特徴ベクトルを考慮した新たな特徴ベクトルを得る．Soft Attention は，各時刻の特徴ベクトルから Attention と呼ばれる重みを推定した後，全時刻の特徴ベクトルから重み付き平均を得る仕組みである [26]．入力の系列データに関して，LSTM によって得られる特徴ベクトルの系列を $\{\mathbf{h}_\tau^1, \mathbf{h}_\tau^2, \cdots, \mathbf{h}_\tau^T\}$ としたとき，Soft Attention によって重み付き平均したベクトル $\mathbf{h}_a$ を次のように計算して得る．

$$\mathbf{h}_a = \sum_{t=1}^{T} \alpha^t \mathbf{h}_\tau^t, \tag{2}$$

$$\alpha^t = \frac{\exp(\mathbf{u}_t^{\mathrm{T}} \mathbf{u}_a)}{\sum_{i=1}^{T} \exp(\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_a)}, \tag{3}$$

$$\mathbf{u}_t = \tanh(\mathbf{W}_a \mathbf{h}_\tau^t + \mathbf{b}_a). \tag{4}$$

ここで，$\mathbf{W}_a \in \mathbb{R}^{U \times U}, \mathbf{b}_a \in \mathbb{R}^U, \mathbf{u}_a \in \mathbb{R}^U$ は Soft Attention を計算するためのモデルパラメータである．

4.3 Grid Embedding Layer

グリッド空間への特徴埋め込み．前方映像中にどの領域にどのような物体があったかの特徴を得るために，物体検出で得られた領域情報を用いて，特徴変換を行う．物体検出の領域情報は，画像に適切な説明文を自動的にアノテーションする画像キャプション生成タスク [25] や，また時系列データを対象としたものではイベントデータからの前方映像中の事故予測 [3] などのタスクで活用されている．これらの研究では検出された物体の名称を識別できる ID と領域情報を各時刻で入力し，深層学習，特に Soft Attention によって検出物体を座標と共に適切な特徴へと変換している．この手法は物体の画像中の位置関係を考慮できる一方で，入力の各時刻で Soft Attention を計算するため計算コストが高い．本論文で取り組むヒヤリハットの発生対象分類においては，入力の各時刻で出力を得る必要はなく，前方映像中の物体間の位置関係よりも，前方映像に写る物体と自車との距離を考慮することが重要であると考えられる．

そこで本論文では，適当な縦と横のグリッド数 G_h, G_w と物体種類数 V からなる行列 $\mathbf{G} \in \mathbb{R}^{G_h \times G_w \times V}$ を用意し，物体検出結果をグリッド空間に埋め込むことによって，物体検出の領域情報を考慮した特徴行列を得る．特徴行列 $\mathbf{G}$ の生成手順を**アルゴリズム 1** に示す．

$\mathbf{G}$ の生成では，入力に物体検出結果の系列 $\mathbf{O} = \{\{\mathbf{o}_n^1\}_{n=1}^{N_1}, \{\mathbf{o}_n^2\}_{n=1}^{N_2}, \cdots, \{\mathbf{o}_n^T\}_{n=1}^{N_T}\}$ と物体の領域情報の系列 $\mathbf{B} = \{\{\mathbf{b}_n^1\}_{n=1}^{N_1}, \{\mathbf{b}_n^2\}_{n=1}^{N_2}, \cdots, \{\mathbf{b}_n^T\}_{n=1}^{N_T}\}$，画像の縦幅 H と横幅 W，縦と横のグリッド数 G_h, G_w，物体の種類数 V，全フレーム数 T を用いる．出力する行列 $\mathbf{G}$ を初期化した後，画像中の設定したグリッド数に対するステップ幅 S_h, S_w を計算する．時刻 t，検出物体 n に関して，その領域情報が画像中のどのグリッドに含まれているかを得るため，左端，上端，右端，下端インデックス $left$，top，$right$，$bottom$ を計算する．また，当該グリッドに埋め込むスコアとして，自車と物体との距離を考慮することを目的に，領域情報の画像サイズに対する面積比 r を計算する．これは，領域情報の面積が大きいほど自車に対して物体は近い距離にあり，また近い距離にある物体ほどヒヤリハット発生対象の要因となっていると考え，その特性を考慮したスコアをグリッド空間に埋め込むためである．そして，対象グリッドについて，縦方向を top から $bottom$，横方向を $left$ から $right$ の範囲で走査し，該当するグリッドに対してスコア r を $\mathbf{g}_{i,j}$ に加算する．

Soft Attention を用いた特徴変換．このようにして得られたグリッド特徴 $\mathbf{g}_{i,j}$ について，ヒヤリハット発生対象分類において重要なグリッドはデータによって異なると考えられる．例えば，**図 1** の左例に示す実際の前方映像のように，ドライブレコーダの設置位置によっては車のボンネットが画像中に大きく写り込むケースがあり，ボンネットが写っている領域には，ヒヤリハット発生対象の要因となる物体が写らないことから，このグリッド領域は発生対象分類に強く寄与しない．ボンネットの領域はドライブレコーダの設置位置によって変動することから，各グリッドの重要度を固定したモデルパラメータとして保持すること

Algorithm 1 Grid embedding

Input: $\mathbf{O}, \mathbf{B}, H, W, G_h, G_w, V, T$
Output: $\mathbf{G}$
Initialize: $\mathbf{G} \in \mathbb{R}^{G_h \times G_w \times V} \leftarrow 0$
$S_h \leftarrow \frac{H}{G_h}$
$S_w \leftarrow \frac{W}{G_w}$
for each time t in 1 to T **do**
  for each detected object n in 1 to N_t **do**
    $left \leftarrow \lceil \frac{b_n^t.left}{S_w} \rceil$
    $top \leftarrow \lceil \frac{b_n^t.top}{S_h} \rceil$
    $right \leftarrow \lceil \frac{b_n^t.right}{S_w} \rceil$
    $bottom \leftarrow \lceil \frac{b_n^t.bottom}{S_h} \rceil$
    $height \leftarrow b_n^t.top - b_n^t.bottom$
    $width \leftarrow b_n^t.right - b_n^t.left$
    $r \leftarrow \frac{height \times width}{H \times W}$
    for each vertical grid index i in top to $bottom$ **do**
      for each horizontal grid index j in $left$ to $right$ **do**
        $\mathbf{g}_{i,j} \leftarrow \mathbf{g}_{i,j} + \mathbf{o}_n^t \cdot r$
      end for
    end for
  end for
end for

は適切でない．

そこで，本論文ではグリッド特徴を以下の Soft Attention によって重み付き平均した特徴ベクトル $\mathbf{h}_g$ を得る．

$$\mathbf{h}_g = \sum_{i=1}^{G_h} \sum_{j=1}^{G_w} \alpha_{i,j} \mathbf{g}_{i,j}, \tag{5}$$

$$\alpha_{i,j} = \frac{\exp(\mathbf{u}_{i,j}^{\mathrm{T}} \mathbf{u}_g)}{\sum_{i=1}^{G_h} \sum_{j=1}^{G_w} \exp(\mathbf{u}_{i,j}^{\mathrm{T}} \mathbf{u}_g)}, \tag{6}$$

$$\mathbf{u}_{i,j} = \tanh(\mathbf{W}_g \mathbf{g}_{i,j} + \mathbf{b}_g). \tag{7}$$

ここで，$\mathbf{W}_g \in \mathbb{R}^{V \times U}, \mathbf{b}_g \in \mathbb{R}^U, \mathbf{u}_g \in \mathbb{R}^U$ は上記の Soft Attention を計算するためのモデルパラメータである．各グリッドに埋め込まれている物体検出結果に対して算出したスコアから動的に重要度 $\alpha_{i,j}$ が算出され，その重要度に基づいて重み付き平均された $\mathbf{h}_g$ を得ることが可能となる．

4.4 Multi-task Layer

4.2 節と **4.3 節**で得られた 2 種類の特徴ベクトル $\mathbf{h}_a$ と $\mathbf{h}_g$ を結合し，特徴ベクトル $\mathbf{h}_{ag} = [\mathbf{h}_a; \mathbf{h}_g]$ を得る．この特徴ベクトルに基づき，ヒヤリハット発生対象を推定する．ここで，本論文ではメインタスクだけでなく，そこから抽出されたサブタスクを推定し，複数のタスクの推定誤差を組み合わせたマルチタスク学習によりニューラルネットワークを最適化する．深層学習におけるマルチタスク学習は，メインタスクに関連するサブタスクを同一のニューラルネットワークで最適化することにより，汎化性能の高いモデルパラメータを学習できることが知られている [2], [4]．本論文では，メインタスクであるヒヤリハットの発生対象分類を 2 つの簡単な分類タスクに切り分けサブタスクとして設定し，3 つの分類タスクからなる目的関数を最適化することで，汎化性能の高いモデルパラメータの学習を目指す．また，簡単なサブタスクの推定結果をメインタスクであるヒヤリハット発生対象分類に活用することで，より高精度に推定ができると考えられる．

1 つ目のサブタスクとして，各々のデータがヒヤリハットか否かを分類する．ここでは，$\mathbf{h}_{ag}$ を FC 及び Sigmoid 関数を通じて，推定結果を表すスカラ値 $y_b \in [0, 1]$ に変換し，サブタスクの正解ラベル t_b との交差エントロピーによって誤差 L_b を次のように計算する．

$$L_b = -\sum_d \{t_b \log y_b + (1 - t_b) \log (1 - y_b)\}. \tag{8}$$

ここで，d はデータを走査するためのインデックスであり，各々の正解ラベルと推定結果に付随するが本論文では省略して記述する．

2 つ目のサブタスクとして，非ヒヤリハットを除くヒヤリハット発生対象を分類する．ここでは，$\mathbf{h}_{ag}$ を FC 及び Softmax 関数を通じて，$C - 1$ 個の次元を持つ推定結果を表すベクトル $\mathbf{y}_c$ に変換し，1-of-K 表現にしたサブタスクの正解ラベル $\mathbf{t}_c$ との交差エントロピーによって誤差 L_c を次のように計算する．

$$L_c = -\sum_d \sum_{k=1}^{C-1} t_{c,k} \log y_{c,k}. \tag{9}$$

これら 2 つのサブタスクで得られたヒヤリハットか否かの推定結果 y_b と，非ヒヤリハットを除いたときのヒヤリハット発生対象の推定結果 $\mathbf{y}_c$ を共に考慮するため，これらの推定結果を $\mathbf{h}_{ag}$ に連結した新たな特徴ベクトル $\mathbf{h}' = [\mathbf{h}_{ag}; y_b; \mathbf{y}_c]$ を得る．これにより，非ヒヤリハットであるか否か，また非ヒヤリハットを除いた場合にヒヤリハット発生対象はいずれかというサブタスクの推定結果を考慮できる．この特徴ベクトル $\mathbf{h}'$ を FC 及び Softmax 関数を通じて，C 個の次元を持つ推定結果を表すベクトル $\mathbf{y}_a$ に変換し，1-of-K 符号化したメインタスクの正解ラベル $\mathbf{t}_a$ との交差エントロピーによって誤差 L_a を次のように計算する．

$$L_a = -\sum_d \sum_{k=1}^{C} t_{a,k} \log y_{a,k}. \tag{10}$$

以上の 3 種類の誤差を合計することによって，ネットワーク全体の目的関数を得る．

$$L = L_a + \beta \cdot (L_b + L_c). \tag{11}$$

L を誤差逆伝播法を用いて最小化することによって，ネットワークを定期化する．ここで，β は追加した 2 つのサブタスクにおける誤差を，どの程度考慮するかを調整するた

めのハイパーパラメータである．

提案手法におけるメインタスクの推定ラベルは，推定結果 $\mathbf{y}_a$ から最大値を持つインデックスを抽出することによって得られる．

5. 評価実験

本章では，実際のイベントデータを用いて定量的に提案手法の有効性を評価する．その評価結果を通じて **1** 章で述べた課題を解決できているかを検証した後，定性評価によって提案手法の実際の推定結果を確認する．

5.1 データセットとパラメータ設定

実験で用いるデータセットには，東京農工大学スマートモビリティ研究拠点の提供する「ヒヤリハットデータベース」を利用する[*3]．ヒヤリハットデータベースは，日本国内のタクシーに設置されたドライブレコーダで記録されたデータから構築されている．ヒヤリハットデータベースでは，衝撃が加わったか否かを判定するドライブレコーダのトリガとして，前後加速度が 0.45G 以上を設定しており，各イベントデータはそのトリガの時刻を起点に前 10 秒，後 5 秒の約 15 秒の前方映像とセンサ系列からなる．また，各イベントデータは訓練された人間によるヒヤリハットレベル（高，中，低，反応）と非ヒヤリハットのラベルに加えて，ヒヤリハット発生対象（車，自転車，バイク，歩行者，単独，その他）のラベルも付与されている．この内，本論文ではヒヤリハットデータとしてヒヤリハットレベルが「中」であるイベントデータについて，発生対象ラベル {車，自転車，バイク，歩行者，単独} からそれぞれ 300 件，また非ヒヤリハットのイベントデータ 700 件をランダムに抽出し，これら合計 2,200 件を対象に実験を行うすなわち，ラベルの種類数は $C = 6$ である．「単独」ラベルはヒヤリハットであるが，その事由が道路のはみ出しや壁への激突の間際，交通違反（信号無視，一時停止不履行）など，自車をその原因とするイベントデータを表す．それぞれのイベントデータは映像，センサ共に 30fps で記録されており，450 の総フレームからなる．本論文では，15 フレーム間隔でデータ点をサンプリングし，$T = 30$ のフレームをそれぞれのイベントデータから抽出した．画像データは RGB 形式で 640×400 $(W = 640, H = 400)$ の解像度で記録されており，YOLO9000 による物体検出は元画像に対して直接処理し[*4]，Pre-training 済みの GoogLeNet による画像特徴抽出は 224×224 に線形変換した後に行った[*5]．センサデータは GPS，ウィンカー操作，ブレーキ操作等ある中から，本論文では予備実験を参考に，前後加速度（GS_Y），横加速度（GS_X），スピード（Speed）の 3 種類を使用した．実験では，用意したデータの内 70%の 1,540 件を訓練データとし，30%の 660 件を評価データとした．各ラベルごとの訓練データと評価データの内訳を**表 2** に示す．また，本論文で提案するマルチタスク学習における各ラベルと追加正解ラベルとの対応関係を同表に示す．以下，ラベル名を {(非ヒヤリハット, No near miss), (車, Car), (自転車, Bicycle), (バイク, Motorcycle), (歩行者, Pedestrian), (単独, Self)} とする．

表 2 使用したデータセットの内訳と各正解ラベルとの対応関係

	正解ラベル			データ件数		
Label	t_a	t_b	t_c	**Train**	**Test**	**Total**
No near miss	1	0	-	504	196	700
Car	2	1	1	212	88	300
Bicycle	3	1	2	206	94	300
Motorcycle	4	1	3	205	95	300
Pedestrian	5	1	4	208	92	300
Self	6	1	5	206	94	300
Total				1,540	660	2,200

提案手法における深層学習部分のパラメータは，各 Fully-Connected Layer のユニット数を $U = 256$ とし，Fully-Connected Layer 後のベクトルを ReLu 関数 [18] で非線形変換し，過学習を抑制する技術として知られるドロップアウト [21] を $p = 0.9$ に設定して入れた．ネットワークの学習では，誤差逆伝搬法で求められた目的関数の勾配に基づいて，Adam [13] によって最適化する[*6]．このとき，mini-batch のサイズは 50 に固定し，誤差逆伝播の回数（Epoch 数）は 100 にしている．また，提案手法は Python で実装し，特にネットワーク構造の定義とその学習は Chainer[*7]で実装した．GoogLeNet に関しては，ImageNet と Places365 で学習された Caffe[10] によるモデルを用い[*8]，出力層のパラメータのみ Fine-tuning によって更新した．

5.2 評価尺度

ヒヤリハットの発生対象分類タスクにおいて，提案手法の有効性を議論するため，推定結果の正解率（Accuracy）に加えて，推定結果がどれだけ正解しているかという正確性と，推定結果が全ての正解のうちどれだけ網羅しているかという網羅性の，2 つの観点から評価する．本論文では，正確性を適合率（Precision），網羅性を再現率（Recall），またこれらの評価指標の調和平均である F 値（F1-score）に

[*3] http://web.tuat.ac.jp/~smrc/drcenter.html

[*4] 実験対象画像に対する物体検出の結果，物体種類数は $V = 69$ となった．中には，1 度しか検出されなかった物体も複数存在したが，出現頻度の低い物体の除去などは一切行っていない．同様に，誤って検出された物体の除去も行っていない．

[*5] 前方映像のカメラの設定位置によっては，ボンネットが大きく写り込むものや，映像が傾いているものもあるが，画像に対する編集操作は線形変換を除いて一切行っていない．

[*6] 最適化のハイパーパラメータは，$\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$ とした

[*7] https://chainer.org

[*8] いずれも，Web 上で公開されているモデルを使用した．
https://github.com/BVLC/caffe/tree/master/models
https://github.com/CSAILVision/places365

よって提案手法の分類性能を評価する [16]．これらの評価指標を計算するため，評価データの集合と推定結果の集合に関して，次のような 4 種類のパラメータを定義する．

TP: 正例に対して推定結果が正であったデータの件数
FP: 負例に対して推定結果が正であったデータの件数
FN: 正例に対して推定結果が負であったデータの件数
TN: 負例に対して推定結果が負であったデータの件数

このとき，Accuracy，Precision，Recall，F1-score は次のように計算される．

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \tag{12}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{14}$$

$$\text{F1}-\text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{15}$$

5.3 実験結果

本節では，実験結果の分析を通じて，**1 章**で課題として挙げた 2 項目が，提案したアプローチで解決できているか否かを検証する．ここでは，4 つの検証事項を提示する．検証 1 と検証 2 は課題 1 の解決の検証，検証 3 は課題 2 の解決の検証に対応している．これに加えて，検証 4 は本論文で提案した 3 つの構成要素を組み合わせることに対する有効性を検証する．

検証 1. Temporal Encoding Layer において，情報源（物体検出結果）を増やすことで分類性能を向上できるか？

ここでは，情報源を組み合わせて時系列の遷移を特徴変換する Temporal Encoding Layer において，その入力する情報源（前方映像，センサ，物体検出結果）を追加することで，ヒヤリハット発生対象の分類性能が向上するか検証する．**表 3** に，Temporal Encoding Layer における，入力する情報源（前方映像，センサ，物体検出結果）の使用の有無に対するヒヤリハット発生対象の分類性能を示す．ここでは，前方映像を Vid，センサを Sen，物体検出結果を Obj と表記している．情報源の追加による分類性能の向上に着目するため，Grid Embedding Layer と Multi-task Layer を除いて実験した．情報源を 1 種類のみにしたケースでは，物体検出結果を使用した手法，情報源を 2 種類にしたケースでは，センサと物体検出結果を組み合わせた手法が，いずれの評価値でも最も高い分類性能を示した．このことから，ヒヤリハット発生対象分類においては，物体検出結果を情報源として使用することの有効性を確認できる．また，前方映像とセンサを使用した手法は，著者らの先行研究 [33] におけるヒヤリハット検出手法に基づくものであるが，この手法は物体検出結果のみを用いる手法に比べて評価値が低いことから，ヒヤリハット発生対象分類には有効に機能しないことが考えられる．このケースを除けば，情報源を追加することでそれぞれの評価値は向上していることから，ヒヤリハット発生対象の分類性能は向上できると示唆される．

表 3 Temporal Encoding Layer における入力する情報源の有無に対する分類性能の比較

Vid	Sen	Obj	Acc	Prec	Rec	F1
✓			41.44	41.31	42.34	41.46
	✓		51.41	49.00	49.77	48.97
		✓	57.48	57.21	56.90	56.20
✓	✓		53.25	52.06	53.11	52.37
✓		✓	61.70	60.58	60.70	60.44
	✓	✓	62.19	61.60	61.91	61.42
✓	✓	✓	**63.38**	**62.77**	**63.13**	**62.85**

検証 2. Grid Embedding Layer は分類性能を向上できるか？

ここでは，物体検出結果に関してその物体の出現位置や種類を考慮して情報をグリッド空間に埋め込み，その後深層学習を用いて特徴変換する Grid Embedding Layer の有効性を検証するため，グリッドサイズに関して，3 種類のグリッドサイズ (G_h, G_w) で分類性能の評価をした．その結果を**表 4** に示す．分類性能がどの程度向上するかを確認するために，Temporal Encoding Layer で全ての情報源を使用し，また Grid Embedding Layer を除いた手法を $(G_h, G_w) = (-, -)$ として同表に加えている．グリッドサイズの選び方は，元画像の解像度 $H = 400$，$W = 640$ のサイズ比 5:8 をもとに，可能な限りサイズ比が崩れないような値を選んだ．また，グリッドサイズに対する分類性能の変化に注目するため，ここでは Multi-task Layer を除いて実験した．実験の結果，いずれのグリッドサイズにおいても，Grid Embedding Layer を使用しない手法に比べて高い評価値を示している．このことから，Grid Embedding Layer はヒヤリハット発生対象の分類性能を向上できると考えられる．また，グリッドサイズを大きく（格子を細かく）していくと，それぞれの評価値も高くなっている．これは，グリッドサイズを大きくしていくことによって，各物体の領域情報をより詳細にグリッド空間に埋め込むことができるためであると考えられる．一方で，グリッドサイズを大きくしていくと，Grid Embedding Layer における Soft attention の計算量も $O(G_h \cdot G_w)$ で大きくなっていくことと，今回実験した最大のグリッドサイズ $(G_h, G_w) = (16, 20)$ に比べて大きく F1-score が低下していないことから，$(G_h, G_w) = (8, 10)$ をこの後の実験では使用した．

表 4　Grid Embedding Layer におけるグリッドサイズ (G_h, G_w) と分類性能の比較

G_h	G_w	Accuracy	Precision	Recall	F1-score
-	-	63.38	62.77	63.13	62.85
5	8	63.40	64.94	64.95	64.78
8	10	65.69	65.69	66.31	65.90
16	20	66.33	66.10	66.31	65.98

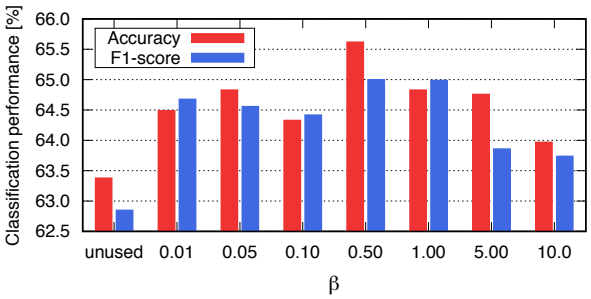

図 4　Multi-task Layer における β に対する分類性能の比較

検証 3.　Multi-task Layer は分類性能を向上できるか？

ここでは，メインタスクの階層構造を利用してサブタスクを抽出し，そのサブタスクの推定結果を活用してメインタスクを推定する Multi-task Layer の有効性を検証するため，サブタスクの誤差を考慮するハイパーパラメータ β に関して，{0.01, 0.05, 0.10, 0.50, 1.00, 5.00 10.00} を設定した際の分類性能の評価をした．ここでは，マルチタスク学習における β の値に対する分類性能の変化に注目するため，Grid Embedding Layer を除いて実験した．その結果を**図 4** に示す．図では，縦軸が Accuracy と F1-score，横軸が β の値である．分類性能がどの程度向上したかを確認するために，Temporal Encoding Layer で全ての情報源を使用し，Multi-task Layer を除いた手法を unused として同図に加えている．実験の結果，今回評価した β に関してはいずれの値においても，Multi-task Layer を使用しない手法に比べて高い評価値を示している．また，β が 0.50, 1.00 で Accuracy と F1-score が他に比べて高くなっており，5.00，10.00 では F1-score が大きく低下している．この結果から，サブタスクの誤差を強く考慮すると，メインタスクの推定精度が低下すると考えられる．評価値が高い結果を示した $\beta = 0.5$ をこの後の実験では使用した．

検証 4. 提案手法における各構成要素を組み合わせることで分類性能を向上できるか？

提案手法における各構成要素を組み合わせることの有効性を検証するため，著者らの既存手法 [33] を Baseline，本論文の提案手法を Proposed として分類性能を評価した．その結果を**表 5** に示す．表では，検証 1，検証 2，検証 3 で得られた結果も，Temp，Temp+Grid，Temp+Ml-task として加えている．いずれの評価値においても，本論文で提案する全ての構成要素を使用した Proposed の評価値が最も高い結果となった．このことから，提案手法における各構成要素を組み合わせることで分類性能を向上できると考えられる．

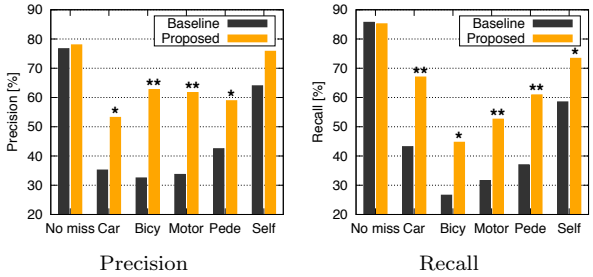

図 5　Proposed と Basline の各クラス別の Precision と Recall

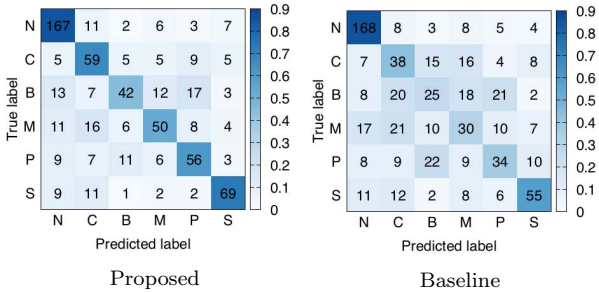

図 6　Proposed と Baseline の分類結果の Confusion matrix

Proposed と Baseline のクラス別の Precision と Recall を図 5 に示す．それぞれの図の縦軸が Precision と Recall である．Proposed と Baseline のクラス別の推定結果と正解を用いて，正解と不正解の件数によってクロス集計表を作成し，χ^2 検定を実施した．その結果，有意水準 5%であるものを「*」，有意水準 1%であるものを「**」として図中に表示している．この結果，適合率では Car，Bicycle，Motorcycle，Pedestrian の 4 クラス，再現率では No near miss を除く 5 クラスで Proposed と Baseline の間に有意な差が認められた．また，全てのクラスを合計した結果においては，適合率と再現率の両方で Proposed と Baseline の間に有意水準 1%で有意な差が認められた．

Proposed と Baseline の分類結果を Confusion Matrix として**図 6** に示す．図では，各ヒヤリハット発生対象のラベルを簡単のため**表 2** の先頭文字を抽出しており，横方向に正解ラベル，縦方向に各手法の推定ラベルを示している．各セルの中の数字は，そのラベルに対してそれぞれの手法が推定した件数を示しており，左上から右下にかけての対角の数字が高いほど，正しく推定していることを表している．また，各セルの色の濃淡は再現率の高さを表している．Baseline に比べて Proposed は，No near miss 以外のラベルで正解の件数が大きく増えており，このことからも，物体認識結果を活用することの有効性を確認できる．

5.4　定性評価

Attention score から見た各ラベルの特性比較．提案

表 5　提案手法の各構成要素が分類性能に与える効果の比較

Method	Video	Sensor	Objects	Grid	Ml-task	Accuracy	Precision	Recall	F1-score
Baseline	✓	✓				53.25	52.06	53.11	52.37
Temp	✓	✓	✓			63.38	62.77	63.13	62.85
Temp+Grid	✓	✓	✓	✓		65.69	65.69	66.31	65.90
Temp+Ml-task	✓	✓	✓		✓	65.62	64.69	65.55	65.00
Proposed	✓	✓	✓	✓	✓	**67.19**	**67.19**	**67.22**	**67.20**

手法では，Temporal Encoding Layer と Grid Embedding Layer で，時間とグリッド空間に対して Soft Attention を用いている．この Soft Attention がテストデータに対して算出した重み $\alpha^t, \alpha_{i,j}$ を，正解ラベルで平均することによって，Soft Attention が各ラベルに対して重視している時間とグリッド空間を比較できる．

時間に対して算出した α^t をラベル別に平均したものを**図 7** に示す．図では縦軸が平均した Attention score，横軸がフレーム数である．ドライブレコーダのトリガー時刻と同じフレーム数は 20 付近である．ヒヤリハット発生対象ラベルでは，20 付近からスコアが高くなっており，特に Car，Bicycle，Motorcycle，Pedestrian は 25 の前でピークになっていることがわかる．Self は最終の 30 フレームに向けてスコアが高くなっていることがわかる．No near miss は 20 の手前からスコアやや高くなり，20 を超えたあたりにピークがある．このように，Self と No near miss は Attention score に他のラベルに比べて明らかに異なる特性を持つが，Car，Bicycle，Motorcycle，Pedestrian は似た傾向を持つことがわかる．これは，先に述べた Confusion matrix による分析で，Bicycle，Motorcycle，Pedestrian の 3 ラベルでは，分類誤りが他のラベルに比べて多い結果となった一要因として示唆される．

グリッド空間に対して算出した $\alpha_{i,j}$ をラベル別に平均したものを**図 8** に示す．図では各セルの濃淡が平均した Attention score の高さを表している．全てのラベルで右側のセルに比べて左側のセルのスコアが高いのは，車が左側車線を走行し，ヒヤリハット発生対象分類に寄与する物体も同車線上に現れるためであると考えられる．セルの中央下部のスコアが低いのは，その付近に自車のボンネット付近が現れるためであると考えられる．Pedestrian は他のラベルに比べて，横グリッドの中央のスコアが高いことから，その付近に歩行者が頻繁に写ることを示唆している．他のラベル間では，縦グリッドの上のスコアが高いか中央のスコアが高いかで違いがあり，これはヒヤリハット発生対象によって自車に対して遠くに写る（Car，Bicycle）か近くに写る（Motorcycle，Self）かに違いがあり，このことを考慮したスコアが得られているためであると考えられる．

実際の分類結果に対する定性的分析．最後に，提案手法が実際にイベントデータに対してヒヤリハット発生対象を分類した結果を，前方映像のいくつかの画像とセンサ

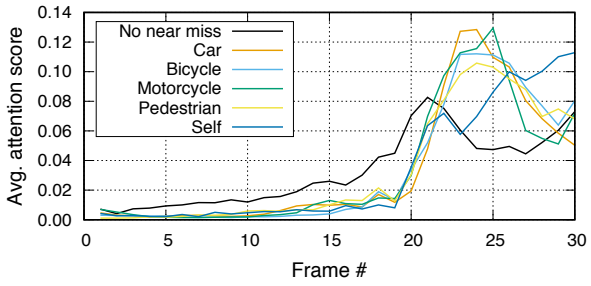

図 7　時間方向の特徴変換に対する Attention の各クラス別の平均

の値と共に確認する．**図 9** は，正解のヒヤリハット発生対象が Motorcycle である例である．Proposed は正しく分類し，Baseline は Pedestrian と誤って分類した．この例では，交差点でバイクが左から右に横切る際に，自車は交差点前で一旦停止し（$t \leq 15$），再び発進したタイミングで（$16 \leq t \leq 19$），左から交差点を横切るバイクを見つけてブレーキで急停止し（$t = 20$），そのバイクが横切るまで停止し続けている（$21 \leq t$）．この例では，実際にヒヤリハット発生対象が前方映像中に出現するのは，ヒヤリハット発生時刻 $t = 20$ ではなく，その少し後 $t = 25$ である．このことから，ヒヤリハット発生対象分類タスクでは，ヒヤリハット発生時刻付近 $t = 20$ のみを分析するだけでなく，時系列データの全体を分析する必要があると考えられる．このことは，**図 7** における Soft Attention のスコアからも示唆できる．また Proposed は，物体検出処理によって得られた「バイク」が埋め込まれている領域や，その物体情報を考慮することができたために，正しく分類できたと考えられる．

図 10 は，正解のヒヤリハット発生対象が Pedestrian である例である．実際，$t = 25$ の画像中央に人物が横断歩道を横切っていることがわかる．しかし，Proposed は誤って Self に，Baseline は誤って Car に分類した．この例は，ヒヤリハット発生対象が複合していると考えられる．その理由として，この車は大きい道路に出る際に（$t < 20$），速度が 0 になっていなことから一時停止しておらず，その結果として横断歩道を横切る歩行者を見つけて急停止し，ヒヤリハットとなっている．すなわち，交通違反に該当する Self を要因として，歩行者に該当する Pedestrian に分類されている．このことから，ヒヤリハット発生対象として Self も間接的には正しいラベルであると考えられ，Self と Pedestrian の両方をラベルとして出力することが望まし

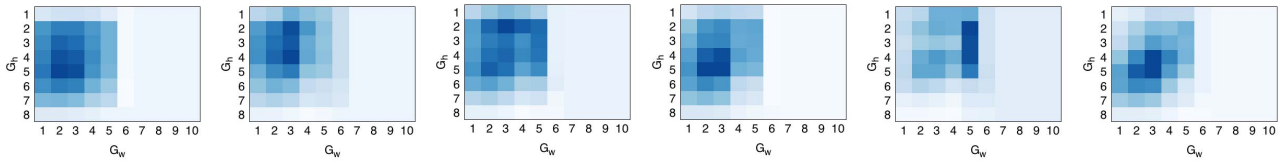

図 8　グリッド空間に対する Soft Attention の各クラス別の平均

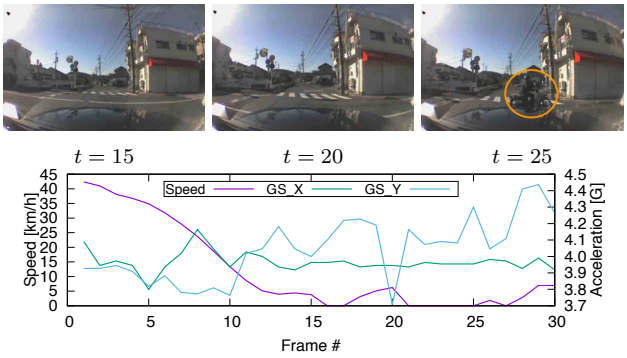

図 9　正解のヒヤリハット発生対象が Motorcycle であった例．Proposed は正しく分類し，Baseline は Pedestrian に誤って分類した．

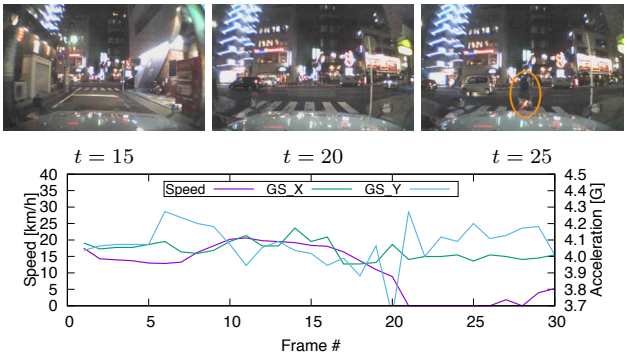

図 10　正解のヒヤリハット発生対象が Pedestrian であった例．Proposed は Self に，Baseline は Car に誤って分類した．

い．この要件を満たすために，現在の 1 つのラベルを出力する方式から，複数のラベルを出力するマルチラベル分類への発展を検討している．

6. おわりに

本論文では，ドライブレコーダで記録されたイベントデータに対してヒヤリハットの発生対象をラベリングすることを目的に，映像とセンサ，物体認識結果を複合的に活用した分類手法を提案した．提案手法では，時系列データの特徴変換（Temporal Encoding Layer），物体認識結果のグリッド空間への特徴埋め込み（Grid Embedding Layer），メインタスクから抽出したサブタスクを用いたマルチタスク学習（Multi-task Layer）という大きく 3 つの要素から構成され，評価実験においてそれぞれの有効性を確認した．それらを組み合わせたときに，より高精度にヒヤリハット発生対象を推定できることを明らかにした．

今後は，より少ない訓練データを対象としたケースや，少ない入力フレーム数を対象としたケースでも高い精度が実現できるように，半教師あり学習への拡張や目的関数の再設計を検討している．

謝辞

本研究は，国立大学法人東京農工大学スマートモビリティ研究拠点の提供する「ヒヤリハットデータベース」を利用した．ここに記して謝意を示す．

参考文献

[1] Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J. and Zieba, K.: End to End Learning for Self-Driving Cars, *CoRR*, Vol. abs/1604.07316 (online), available from ⟨http://arxiv.org/abs/1604.07316⟩ (2016).

[2] Bordes, A., Glorot, X., Weston, J. and Bengio, Y.: Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing, *In AISTATS*, pp. 127–135 (2012).

[3] Chan, F.-H., Chen, Y.-T., Xiang, Y. and Sun, M.: Anticipating Accidents in Dashcam Videos, *In ACCV*, pp. 136–153 (2017).

[4] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P.: Natural Language Processing (Almost) from Scratch, *Journal of Machine Learning Research*, Vol. 12, pp. 2493–2537 (2011).

[5] Driving Safety Promotion Center Honda Motor Co. Ltd.: Honda Driving Safety Promotion Activities 2016, , available from ⟨`http://www.honda.co.jp/safetyinfo/global/safetyinfo_2016_E.pdf`⟩ (accessed 2017-10-30).

[6] Fukushima, K.: Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biological Cybernetics*, Vol. 36, pp. 193–202 (1980).

[7] Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, *Neural computation*, Vol. 9, pp. 1735–80 (1997).

[8] Jain, A., Koppula, H. S., Raghavan, B., Soh, S. and Saxena, A.: Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models, *In ICCV*, pp. 3182–3190 (2015).

[9] Jain, A., Koppula, H. S., Raghavan, B., Soh, S. and Saxena, A.: Recurrent Neural Networks for Driver Activity Anticipation via Sensory-Fusion Architecture, *In ICRA*, pp. 3118–3125 (2016).

[10] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell,

T.: Caffe: Convolutional Architecture for Fast Feature Embedding, *In MM*, pp. 675–678 (2014).

[11] Ke, R., Lutin, J., Spears, J. and Wang, Y.: A Cost-effective Framework for Automated Vehicle-pedestrian Near-miss Detection through Onboard Monocular Vision, *In CVPR workshop*, pp. 25–32 (2017).

[12] Kim, J. and Canny, J. F.: Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention, *CoRR*, Vol. abs/1703.10631 (online), available from ⟨http://arxiv.org/abs/1703.10631⟩ (2017).

[13] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol. abs/1412.6980 (online), available from ⟨http://arxiv.org/abs/1412.6980⟩ (2014).

[14] Lam, H. T.: A Concise Summary of Spatial Anomalies and Its Application in Efficient Real-time Driving Behaviour Monitoring, *In SIGSPATIAL*, pp. 30:1–30:9 (2016).

[15] Maeda, M., Uetani, T. and Takagi, M.: Development of Drive Recorder (OBVIOUS Recorder) (2006).

[16] Manning, C. D., Raghavan, P. and Schutze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).

[17] Mikolov, T., Karafit, M., Burget, L., Cernock, J. and Khudanpur, S.: Recurrent neural network based language model, *In INTERSPEECH 2010*, Vol. 2, pp. 1045–1048 (2010).

[18] Nair, V. and Hinton, G. E.: Rectified Linear Units Improve Restricted Boltzmann Machines, *In ICML*, ICML'10, pp. 807–814 (2010).

[19] Redmon, J. and Farhadi, A.: YOLO9000: Better, Faster, Stronger, *arXiv preprint arXiv:1612.08242* (2016).

[20] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *IJCV*, Vol. 115, No. 3, pp. 211–252 (2015).

[21] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, Vol. 15, pp. 1929–1958 (2014).

[22] Suzuki, T., Aoki, Y. and Kataoka, H.: Pedestrian Near-Miss Analysis on Vehicle-Mounted Driving Recorders, *In MVA*, pp. 416–419 (2017).

[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *In CVPR*, pp. 1–9 (2015).

[24] Xu, H., Gao, Y., Yu, F. and Darrell, T.: End-to-end Learning of Driving Models from Large-scale Video Datasets, *In CVPR*, pp. 2174–2182 (2017).

[25] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, *In ICML*, pp. 2048–2057 (2015).

[26] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J. and Hovy, E. H.: Hierarchical Attention Networks for Document Classification, *HLT-NAACL* (2016).

[27] Yokoyama, D. and Toyoda, M.: Do Drivers' Behaviors Reflect Their Past Driving Histories? - Large Scale Examination of Vehicle Recorder Data, *In the BigData Congress 2016*, pp. 361–368 (2016).

[28] Yokoyama, D., Toyoda, M. and Kitsuregawa, M.: Understanding Drivers' Safety by Fusing Large Scale Vehicle Recorder Dataset and Heterogeneous Circumstantial Data, *In PAKDD*, pp. 734–746 (2017).

[29] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. and Torralba, A.: Places: A 10 million Image Database for Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[30] スマートモビリティ研究拠点：ドライブレコーダデータセンター，東京農工大学大学院工学研究院（オンライン），入手先 ⟨http://web.tuat.ac.jp/~smrc/drcenter.html⟩（参照 2017-6-12）.

[31] 菊池理人，日景由華，御室哲志：ドライブレコーダデータからの自動分別の試み，計測自動制御学会東北支部第 290 回研究集会 (2014).

[32] 警察庁交通局：ドライブレコーダーの活用について，警察庁（オンライン），入手先 ⟨https://www.npa.go.jp/bureau/traffic/anzen/drive_recorder.html⟩（参照 2017-6-12）.

[33] 山本修平，遠藤結城，戸田浩之：映像とセンサ信号を用いたドライブレコーダデータからのヒヤリハット検出手法，TOD（テクニカルノート），Vol. 10, No. 4, pp. 26–30 (2017).

[34] 森村哲郎，谷澤悠輔，山崎慎也， 井手剛：統計的機械学習を用いたプローブカーデータからのヒヤリハット発生形態の推定，自動車技術会秋季学術講演会 (2011).

[35] 日本カーソリューションズ株式会社：全国の 38 拠点から安全でエコな自動車リースを展開，NTT 技術ジャーナル（オンライン），入手先 ⟨http://www.ntt.co.jp/journal/1110/files/jn201110040.pdf⟩（参照 2018-5-11）.

[36] 豊田正史，横山大作，伊藤正彦：運転状況を考慮したドライブレコーダデータからの潜在リスク交差点検知手法，*DEIM Forum* (2017).

[37] 林彩和菜：2017 年のドライブレコーダーの販売動向 販売台数は前年から 38%増加，GfK Japan（オンライン），入手先 ⟨https://www.gfk.com/fileadmin/user_upload/dyna_content/JP/20180219_drivingrecorders2.pdf⟩（参照 2018-5-8）.