カメラで撮影された歩行中の人物グループから生成された インタラクション動画を用いた会話の活発さ認識 *

我那覇航 ** 尾崎匠 ** 井上路子 ** 西山正志 **

Conversation activity recognition using interaction video sequences acquired from groups of pedestrians

Wataru GANAHA, Takumi OZAKI, Michiko INOUE and Masashi NISHIYAMA

We propose a method for recognizing the conversation activity in a group of pedestrians walking outdoors using a camera video sequence. Various existing methods have been developed for analyzing interaction in conversation, e.g., action recognition of whether people are talking or not. However, a method has yet to be proposed to recognize the conversation activity in a group of pedestrians. In this paper, we design a method to extract features for conversation activity recognition by using body interactions that occurred from pedestrians' conversations. Our method generates an interaction video sequence in a virtual road space using a temporal pose signal and temporal position signal representing body interactions. Our method determines a conversation activity label using the interaction video sequence, which is an informative and visible feature. Experimental results show that our method using an interaction video sequence can recognize the conversation activity more accurately than comparison methods using a temporal pose signal and temporal position signal.

Key words: conversation activity, interaction video sequence, recognition, pedestrian, group

1. はじめに

屋外で歩行中の人物グループを対象とし，そのグループ内のインタラクションを，自動で認識する技術が求められている．歩行中の人物グループにおけるインタラクションには様々なものがあるが，ここでは会話の活発さに注目する．会話の活発さとは，人物グループ内で会話が発生しており，その会話が盛り上がっているかどうかを指す．会話の活発さ認識の応用シーンとして，店舗における通路など，歩行中の人物グループが行き交う場面でのマーケティングが考えられる．その応用シーンの例を図 **1** に示す．店舗と駐車場との間で，会話が盛り上がっている人物グループの件数を行きと帰りで比較することで，来客者が満足しているかどうかを把握できる可能性がある．

屋外で歩行中の人物グループを対象とし，その会話の活発さを認識するため，何を手掛かりとして利用できるかについて考える．最初に考えられる手掛かりとして，互いの発話タイミングや互いの声の抑揚など，音声の時系列変化が挙げられる．ただし本論文では，屋外で歩行中の人物グループを対象としているため，音声をマイクでセンシングすることは難しく，音声そのものを会話の活発さ認識の手掛かりとして利用できない．ここでは，カメラで撮影される人物グループの身体の見え方において，会話の活発さを認識するための手掛かりが何であるかを考える．本論文では，グループ内の各人物が行うジェスチャの時系列変化と，互いの身体向きおよび互いの歩行位置の時系列変化が，活発さ認識の手掛かりになると仮定する．ジェスチャとは発話に伴い身体で生じる動きであり，会話を分析する上で有効であることが知られている [1]．互いの身体向きと互いの歩行位置とはグループに属する人物らの相互関係であり，歩行者グループ行動の分析 [2] やグループ検出の手法 [3] で有効であることが報告されている．本論文では，ジェスチャと互いの身体向きと互いの歩行位置とを合わせて身体インタラクションと呼ぶことにする．

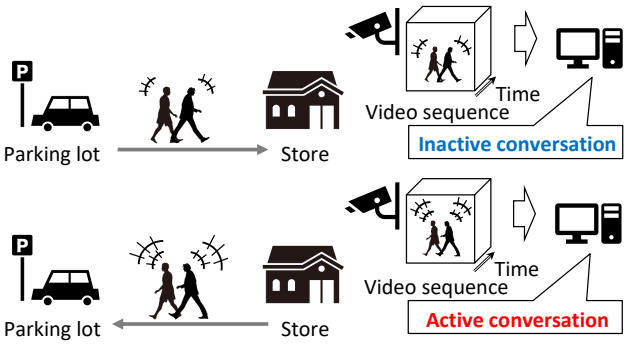


Fig. 1 Example of applications for conversation activity recognition. It is possible to determine whether pedestrian groups are satisfied with their visit by comparing the number of the groups having active conversations between a store and a parking lot on their way to and from a store.

ここで，カメラ動画中の身体インタラクションのみを手掛りとして，会話の活発さを認識する手法を設計することを考える．実際に歩行中の人物グループから会話の活発さを認識できる手法は，我々が調査した限りであるが，これまで提案されていなかった．ここでは，本論文の目的である会話の活発さに特化せず，一緒に歩いているかどうかや，一緒に話しているかどうかなど，人物グループで発生するインタラクションの有無を，カメラ動画を用いて認識する既存手法を以下で紹介する．これらの既存手法は，大きく二つに分けることができる．一つ目は，一緒に歩く複数の人物で構成されるグループの有無を検

* 原稿受付 令和 5 年 5 月 8 日
* 掲載決定 令和 5 年 10 月 31 日
** 鳥取大学大学院工学研究科 (鳥取市湖山町南 4 丁目 101)

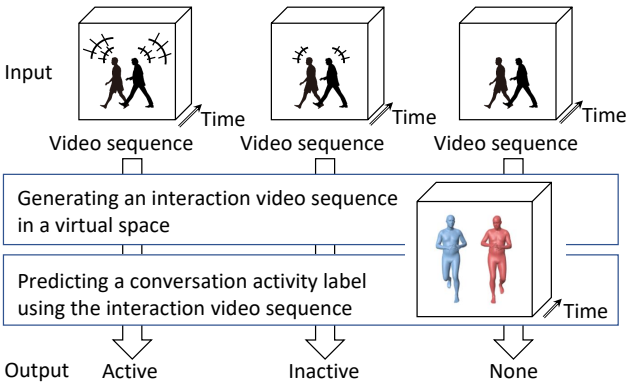


Fig. 2 Overview of our method for conversation activity recognition.

出する手法 [4, 5, 3, 6, 7] である．二つ目は，歩行中の人物グループが検出されたとして，一緒に話しているかどうかについて，会話の有無を認識する手法 [8, 9, 10] である．さらに近年では，歩行中の人物グループの有無を検出すると同時に，そのグループ内での会話の有無を認識する手法 [11] も登場している．ただし，これらの既存手法を適用するとしても，歩行中の人物グループにおける会話について，その会話の活発さを認識することはできなかった．

そこで本論文では，屋外で歩行中の人物グループを対象とし，高い認識精度が得られ，かつ，人間が目視でも確認できるインタラクション動画を特徴量として抽出することで，会話の活発さを認識する手法について述べる．提案手法では，カメラ動画から推定された姿勢の時系列信号，および，歩行位置の時系列信号を用いることで，人物グループ内のインタラクション動画を仮想空間において生成する．このインタラクション動画をクラス分類ネットワークに適用することで，会話が活発である，非活発である，もしくは，会話がないかのラベルを出力する．提案手法の流れを図 **2** に示す．実験結果より，提案手法のインタラクション動画を特徴量として用いることで，人物領域の動画を特徴量として用いる場合や，姿勢の時系列信号と歩行位置の時系列信号とを特徴量として用いる場合に比べて，会話の活発さを精度良く認識できることを確認した．以下では，2. で関連研究を述べ，3. で提案手法について述べ，4. で実験結果について述べる．最後に，6. でまとめる．

2. 関連研究

2.1 会話に関する既存の分析研究

会話に関する分析研究は古くから行われており，特に会議室での会話を想定したものが多く行われている．既存の分析研究 [12, 13, 14, 15] では，会議室での会話をセンシングするため，人物の胸元に取り付けられたマイクと，人物の正面に設置されたカメラとを併用することが一般的である．なお，会議室における会話の分析研究では，着席して大きく動かない人物の会話を，マイクとカメラとを用いてピンポイントでセンシングする手法が採用されている．

ここで，会議室における会話のセンシング手法を，屋外で歩行中の人物グループにおける会話のセンシングにそのまま適用する場合を考える．そのためには，歩行中の人物グループの近くへ，マイクを常に設置する必要がある．例えば，手持ちのスマートフォンを利用することが考えられるが，各人物からオプトインなどの協力を仰ぐ必要があり，屋外で歩行中の人物グループから音声をマイクでセンシングすることは手間がかかると考えられる．次に，カメラを利用する場合を想定すると，センシングにそれほど手間はかからないと考えられる．近年の人物画像認識の技術 [16, 17] の発展により，人物グループの近くにカメラを設置する必要は無いと言える．既設の防犯カメラと既存の人物画像認識の技術とを適切に組み合わせることで，屋外で歩行中の人物グループから，会話中に発生する身体インタラクションを手軽にセンシングできる状況になりつつあると考えられる．本論文では，カメラでセンシングされた身体インタラクションを用いることで，会話の活発さ認識に有効な特徴量をいかに抽出するかについて議論していく．

2.2 グループの有無を検出する既存手法と会話の有無を認識する既存手法

カメラ動画において，複数名で構成されるグループの有無を検出する既存手法 [4, 5, 3, 6, 7] について述べる．既存手法 [4, 5, 6, 7] では，各時刻における画像において，人物の位置を推定し，そこから得られる人物間の距離と，人物の移動速度とを用いてグループの有無を検出している．また，既存手法 [3] では，各時刻の画像における人物の位置に加えて，人物の身体の中でも顔向きを推定することで，グループの有無を検出している．いずれの手法においても，グループの有無を検出するために，人物の位置を推定することが重要となる．提案手法では，既存手法の知見を活用し，画像中の人物位置を推定し，さらに実空間における人物位置へ変換することを狙う．

次に，検出された人物グループの中で，会話の有無を認識する既存手法 [8, 9, 10] について述べる．既存手法 [8, 10] では，会話の有無に加えて，横断，待機，並ぶ，歩くなどの人物のインタラクションの種類を認識している．また，既存手法 [9] では，人物グループ内のインタラクションが，喧嘩などの異常な行動であるかどうかも認識している．さらに既存手法 [11] では，これらのインタラクションの種類の認識に加えて，グループの有無も自動で検出している．これらの既存手法 [8, 9, 10, 11] では，いずれもカメラ動画中の人物の見え方の時系列変化そのものを，会話の有無の認識の手掛かりとして用いている．ただし，動画中の人物の見え方は，照明やカメラ向きなど様々な変動要因で決定される．人物の見え方のみで会話の有無を安定に認識するためには，様々な変動を含む訓練サンプルを大量に準備する必要がある．提案手法では，大量の訓練サンプルを準備することなく人物の見え方変動の影響を抑え，会話中に発生する身体インタラクションを表す特徴量を抽出することで，会話の活発さ認識の精度を高めることを狙う．

3. 会話の活発さ認識のための提案手法

3.1 概要

提案手法では，1. で述べたように，屋外で歩行中の人物グループを対象とし，そのグループ内における会話の活発さを認識するため，身体インタラクションを表現するインタラクション動画を特徴量として，カメラ動画から抽出する．この特徴抽出を設計するため，ここではまず，歩行中の人物グループについて，カメラ動画中での見え方を決定する変動要因を考えていく．それらの変動要因として，以下が挙げられる．

- 人物：姿勢，体型，歩行位置など
- 環境：照明，天候，時間帯など

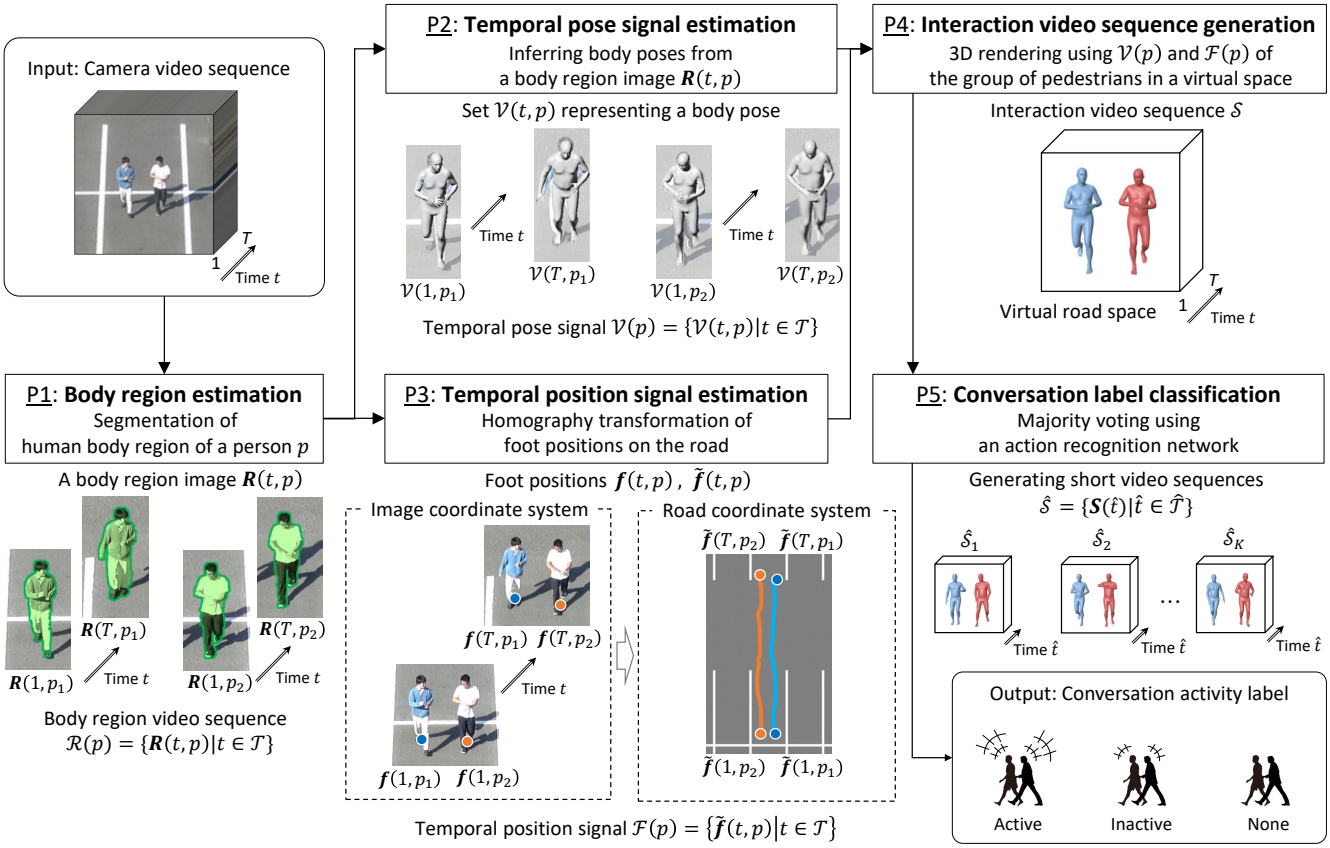


Fig. 3 Overview of our method for conversation activity recognition. We estimate body region images from a camera video sequence in P1 and a temporal pose-position signal for representing the conversation's body interactions in P2 and P3. We generate an interaction video sequence in a virtual road space for extracting an informative and visible feature in P4 and determine a conversation activity label using the interaction video sequence in P5.

- カメラ：カメラ向き，解像度，フレームレートなど

これらの変動要因が複雑に絡み合い，歩行中に会話する人物グループの見え方が変化していく．

本論文では，会話の活発さから発生する身体インタラクションは，上記の変動要因の中で，姿勢と歩行位置とによって特に表現されると仮定する．これら以外の変動要因は，認識精度を低下させると考えられるため，特徴抽出の際に抑えていくことを考える．提案手法ではまず，歩行中の人物グループを撮影したカメラ動画から，姿勢の時系列信号，および，歩行位置の時系列信号を推定する．推定された姿勢の時系列信号と歩行位置の時系列信号とだけを用いて，人物グループのインタラクション動画を仮想空間においてレンダリングし，そのインタラクション動画を特徴量として会話の活発さ認識に利用する．仮想カメラの視点位置を人物グループの正面へ常に固定することで，高い精度が得られる特徴量を設計することを狙う．また，人間が目視でもインタラクションの様子を確認できる特徴量を設計することを狙う．

3.2 提案手法の流れ

提案手法の流れを図 **3** に示し，その流れにおける各手続きについて以下で述べる．

P1. 人物領域の推定

カメラ動画の各時刻において，歩行中の人物の身体を表す領域を推定する．

P2. 姿勢の時系列信号の推定

各時刻における人物領域の見え方から，姿勢の時系列信号を推定する．具体的には，3 次元人体モデルを用いることで，各人物の姿勢のみを表す時系列変化を取り出す．これにより，会話における顔向きや体向きの関係を表現しつつ，認識精度を低下させる見え方の変動要因を抑えることを狙う．

P3. 歩行位置の時系列信号の推定

各時刻における人物領域から足元重心を算出することで，歩行位置の時系列信号を推定する．具体的には，画像上の人物の足元輪郭の重心を推定し，人物が歩行する路面座標系への射影変換を適用することで，路面上での歩行位置を決定する．これにより，会話における位置関係を表す時系列変化を取り出すことを狙う．

P4. インタラクション動画の生成

手続き P2 で推定された姿勢の時系列信号と，手続き P3 で推定された歩行位置の時系列信号とを用いることで，人物グループに属する各人物を仮想空間に配置し，インタラクション動画を三次元レンダリングで生成する．仮想カメラ視点を人物グループの真正面へ常に固定することで，会話の活発さ認識に有効な身体インタラクションを捉えることができ，高い精度が得られる特徴量を設計することを狙う．また，人間が目視でも身体インタラクションを確認できる特徴量を設計することを狙う．

P5. 会話の活発さラベルの分類

手続き P4 で生成されたインタラクション動画について，

会話の活発さを表すラベルを分類ネットワークを用いて決定する．ここでは，会話の活発さラベルを，活発，非活発，および，会話なしの 3 種類とする．一つのインタラクション動画から複数の短動画を生成し，それら短動画を用いて分類ネットワークから候補ラベルを複数出力する．それら候補ラベルから多数決を行うことで，会話の活発さラベルを最終的に決定する．

以下では，各手続きの詳細について述べる．

3.3 人物領域の推定

提案手法の手続き P1 では，カメラ動画から人物領域を推定する．人物 p について，推定された人物領域を含む矩形領域からなる動画 $\mathcal{R}(p)$ を式 (1) で表す．

$$\mathcal{R}(p) = \{\boldsymbol{R}(t,p) \mid t \in \mathcal{T}\} \tag{1}$$

ここで，$\boldsymbol{R}(t,p)$ は人物 p の時刻 t における領域画像，$\mathcal{T}$ は領域画像が撮影された時刻 t を要素として持つ集合とする．集合 $\mathcal{T}$ に属する時刻の総数を T とする．この T は，カメラ動画の視野の中に人物 p が入ってから外れるまでの時間の長さも表している．なお $\boldsymbol{R}(t,p)$ は，人物領域とそれを囲む背景領域との見え方から構成されている．具体的には，各画素が人物領域に属するか背景領域に属するかを表すマスクと，各画素の RGB 値とが格納されている．本論文では，人物領域を推定するため，PHALP [17] の内部から呼び出される Mask R-CNN [16] を各時刻において使用する．なお PHALP とは，次節で述べる姿勢の時系列信号の推定に用いる手法であり，時刻間における人物追跡の処理も同時に内部で実行される．この追跡結果を用いて領域画像 $\boldsymbol{R}(t,p)$ の人物 p を設定する．

3.4 姿勢の時系列信号の推定

提案手法の手続き P2 では，人物 p の領域動画 $\mathcal{R}(p)$ において，会話における顔向きや体向きの関係を表現するため，姿勢の時系列信号を推定する．まず，時刻 t の領域画像 $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$ から人物の姿勢を推定する．ここでは姿勢を，人物の体表上の 3 次元頂点とそれら頂点の隣接関係とからなる集合 $\mathcal{V}(t,p)$ で表す．人物 p の姿勢の時系列信号 $\mathcal{V}(p)$ を式 (2) で表す．

$$\mathcal{V}(p) = \{\mathcal{V}(t,p) \mid t \in \mathcal{T}\} \tag{2}$$

本論文では，姿勢を表す $\mathcal{V}(t,p)$ を推定するため，前節の手続き P1 で述べた PHALP [17] を適用する．この PHALP では，カメラ動画から人物領域を推定し追跡処理を行った上で，3 次元人体モデルの一つである SMPL [18] を用いて，姿勢パラメータと体型パラメータとを表現している．ただし PHALP では，体型は人物間で変化しないと仮定し，姿勢パラメータのみを推定している．姿勢パラメータは，具体的には人体の 23 個の関節点における回転行列と全身における回転行列とで表現されている．推定された SMPL の姿勢パラメータと固定値の体型パラメータとを用いることで，体表上の 6890 個の 3 次元頂点 $\boldsymbol{v}(t,p)$ とそれら頂点の隣接関係とからなる集合 $\mathcal{V}(t,p)$ へ変換する．

姿勢の時系列信号を推定する際，時間方向の外れ値が突然発生することがある．本論文では，3 次元頂点 $\boldsymbol{v}(t,p)$ の時系列信号に Hampel フィルタを適用することで外れ値を検知する．外れ値が存在する時刻の姿勢パラメータを，周辺時刻の値から最近傍法で補間する．

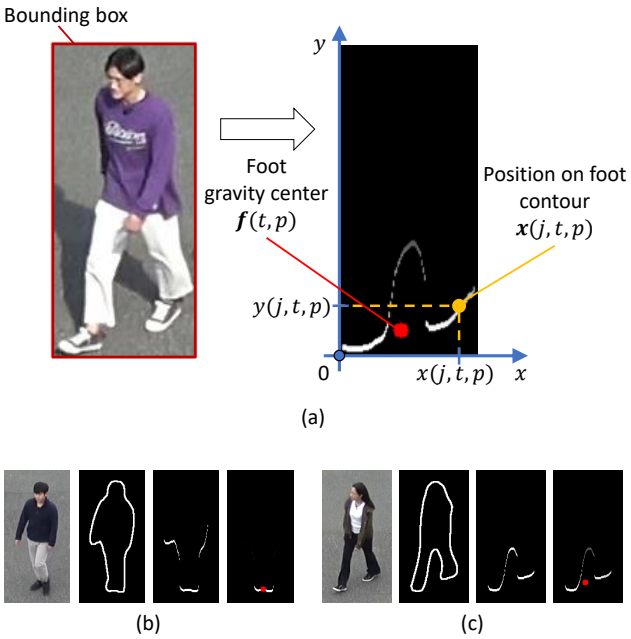


Fig. 4 Parameters used to calculate a foot gravity center $\boldsymbol{f}(t,p)$ of the image coordinate system of time t of person p in P3.

3.5 歩行位置の時系列信号の推定

提案手法の手続き P3 では，人物 p の領域動画 $\mathcal{R}(p)$ において，会話における位置関係を表すため，路面上における歩行位置の時系列信号を推定する．まず，時刻 t の領域画像 $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$ において，画像座標系における足元重心位置 $\boldsymbol{f}(t,p)$ を推定する．次に，画像座標系から実空間の路面座標系への射影変換を適用することで，路面上の人物の歩行位置 $\tilde{\boldsymbol{f}}(t,p)$ を求める．手続き P3 から最終的に出力される歩行位置の時系列信号 $\mathcal{F}(p)$ を式 (3) で表す．

$$\mathcal{F}(p) = \{\tilde{\boldsymbol{f}}(t,p) \mid t \in \mathcal{T}\} \tag{3}$$

ここで，画像座標系における足元重心位置 $\boldsymbol{f}(t,p)$ の算出方法について述べる．この位置を算出する際に用いるパラメータを図 **4**(a) に示す．人物 p の時刻 t の領域画像 $\boldsymbol{R}(t,p)$ において，足元の輪郭上の点の画像座標 $\boldsymbol{x}(j,t,p) = (x(j,t,p), y(j,t,p))$ を取得する．ここで，$\mathcal{J} = \{j\} : j$ は自然数 とし，$\forall j,k \in \mathcal{J} : j < k \Rightarrow x(j,t,p) < x(k,t,p)$ とする．ここでは，人物領域の外接矩形の左下を原点とする．外接矩形の下端 $(x(j,t,p),0)$ から足元輪郭までの距離である成分 $y(j,t,p)$ を用いて，重み $w(j,t,p)$ を式 (4) で求める．

$$w(j,t,p) \sim \mathcal{N}(y(j,t,p)|0,\sigma^2) \tag{4}$$

ここで，$\mathcal{N}(y(j,t)|0,\sigma^2)$ を平均 0，標準偏差 σ の正規分布とする．なお，$w(j,t,p)$ は式 (5) を満たすものとする．

$$\sum_{j\in\mathcal{J}} w(j,t,p) = 1 \tag{5}$$

時刻 t における人物 p の足元重心位置 $\boldsymbol{f}(t,p)$ を式 (6) で求める．

$$\boldsymbol{f}(t,p) = \sum_{j\in\mathcal{J}} w(j,t,p)\boldsymbol{x}(j,t,p) \tag{6}$$

画像座標系での 2 次元足元重心位置 $\boldsymbol{f}(t,p)$ へ，ホモグラフィ行列による射影変換を適用し，路面上での高さを 0 とすること

で，路面座標系での 3 次元歩行位置 $\tilde{\boldsymbol{f}}(t,p)$ を求める．なお，路面の特徴点を用いてホモグラフィ行列を事前に求めておくこととする．

足元輪郭上の点 $\boldsymbol{x}(j,t,p)$ へ与える重み $w(j,t,p)$ を正規分布とする理由について述べる．歩行中の人物は両脚の開閉を連続で繰り返す．図 **4**(b) に脚が閉じている場合の例を示し，(c) に脚が開いている場合の例を示す．提案手法では，足元輪郭の候補点を求めるため，人物領域を囲う閉曲線において，外接矩形の下端から上端に見て最初に出現する点 $\boldsymbol{x}(j,t,p)$ の集合を用いる．脚を閉じている場合の足元輪郭の候補点には，図 **4**(b) 中部に示す通り，主に足元に該当する部分が表れているが，手なども一部表れている．また，脚を開いている場合の足元輪郭の候補点には，図 **4**(c) 中部に示す通り，足元に該当する部分に加えて，股なども表れている．足元に属さない候補点の影響を抑えるため，それらの点は外接矩形の下端から離れていると仮定し，式 (6) において正規分布を用いて小さな重みを与える．

路面上における人物の影や白線などの表記物によって，足元の輪郭が正しく推定されず，歩行位置の時系列信号 $\mathcal{F}(p)$ に外れ値が含まれる場合がある．本論文では，Hampel フィルタを適用することで外れ値を検知し，時間方向の線形補間を用いて歩行位置の補間処理を行う．

3.6 インタラクション動画の生成

提案手法の手続き P4 では，人間が目視でも身体インタラクションを確認できる特徴量を設計するために，姿勢の時系列信号 $\mathcal{V}(p)$ と歩行位置の時系列信号 $\mathcal{F}(p)$ とを用いて，仮想空間に人物を配置し，インタラクション動画 $\mathcal{S}$ を三次元レンダリングで生成する．この仮想空間では，人物グループに属する各人物の姿勢の時系列信号と歩行位置の時系列信号とが，人物の標準体型パラメータを用いて可視化される．人物の標準体型パラメータとは，3.4 で述べた SMPL で準備されている平均的な人物体型を指す．提案手法では，仮想空間における人物の身体インタラクションのみを表現する映像を用いることで，特徴量を設計することを狙う．仮想空間では，あらゆる場所に仮想カメラ視点を置きレンダリングすることが可能であるが，本論文では，会話の活発さ認識の精度向上に寄与すると想定される身体インタラクションを捉えることを狙い，仮想カメラ視点を，人物グループの真正面の定位置へ常にどの時刻においても固定する．

以下ではインタラクション動画 $\mathcal{S}$ の生成方法について述べる．まず，3.4 で求めた人物 p の体表上の 3 次元頂点 $\boldsymbol{v}(t,p) \in \mathcal{V}(t,p) \in \mathcal{V}(p)$ に，3.5 で求めた路面上の歩行位置 $\tilde{\boldsymbol{f}}(t,p) \in \mathcal{F}(p)$ を仮想空間において反映させる．具体的には，仮想空間における路面上の歩行位置が反映された 3 次元頂点を $\tilde{\boldsymbol{v}}(t,p)$ とし，式 (7) で求める．

$$\tilde{\boldsymbol{v}}(t,p) = \boldsymbol{v}(t,p) + \tilde{\boldsymbol{f}}(t,p) \tag{7}$$

集合 $\mathcal{V}(t,p)$ に属する全ての頂点 $\boldsymbol{v}(t,p)$ について $\tilde{\boldsymbol{v}}(t,p)$ を求める．得られた $\tilde{\boldsymbol{v}}(t,p)$ とそれらの隣接関係からなる集合を $\tilde{\mathcal{V}}(t,p)$ とする．仮想空間の路面上における歩行位置が反映された姿勢の時系列信号を $\tilde{\mathcal{V}}(p)$ とし，式 (8) で表す．

$$\tilde{\mathcal{V}}(p) = \{\tilde{\mathcal{V}}(t,p) \mid t \in \mathcal{T}\} \tag{8}$$

次に，一つの人物グループを構成する人物 p を決定するため，各人物の歩行位置 $\tilde{\boldsymbol{f}}(t,p)$ を用いて，人物間の距離を求める．この距離が閾値以下であれば，同じ人物グループに属することとする．人物グループ毎に $\tilde{\mathcal{V}}(p)$ を求めた上で，同一グループに属する人物を仮想空間に配置し，三次元レンダリングを行うことで，インタラクション動画の各時刻の画像 $\boldsymbol{S}(t)$ を生成する．最後に，人物グループ毎のインタラクション動画 $\mathcal{S}$ を式 (9) で表す．

$$\mathcal{S} = \{\boldsymbol{S}(t) \mid t \in \mathcal{T}\} \tag{9}$$

なお，上記で算出された姿勢の時系列信号 $\tilde{\mathcal{V}}(p)$ を仮想空間のレンダリングにそのまま用いた場合，路面に対して人体が不自然に大きく傾いて姿勢パラメータが推定されることがある．提案手法では，路面の垂直方向の軸を除いた残り軸に対する回転角を 0 度と常に設定することで，路面に対する人体の傾きを補正する．

3.7 会話の活発さラベルの分類

提案手法の手続き P5 では，インタラクション動画 $\mathcal{S}$ を用いて会話の活発さを表すラベルを決定するため，行動認識の分野で提案されている既存の分類ネットワークを適用する．本論文では，分類ネットワークとして，三次元畳み込みに基づく C3D [19] を用いる．C3D は，空間方向の畳み込みに加えて，時間方向の畳み込みも同時に実行する．本論文では，インタラクション動画を複数の短動画に分割し，それらの短動画を C3D ネットワークへ入力し，会話の活発さを表すラベルの候補を複数予測する．それらの候補から多数決で最終的なラベルを決定する．なお，会話の活発さラベルを，3.2 の P5 で述べたように，活発，非活発，および，会話なしの 3 種類とする．

会話の活発さを表すラベルを決定する手法の詳細について述べる．C3D の学習時，および，予測時に，一個のインタラクション動画 $\mathcal{S}$ から，初期時刻の異なる短動画 $\hat{\mathcal{S}}$ を K 個生成する．短動画 $\hat{\mathcal{S}}$ を式 (10) で表す．

$$\hat{\mathcal{S}} = \{\boldsymbol{S}(\hat{t}) \mid \hat{t} \in \hat{\mathcal{T}}\} \tag{10}$$

ここで，$\hat{\mathcal{T}}$ は短動画に属する画像 $\boldsymbol{S}(\hat{t})$ の時刻 $\hat{t}$ からなる集合とする．まず，ランダムに初期時刻 $\hat{t}_1$ を決定する．初期時刻 $\hat{t}_1$ から等間隔 I で時刻を進めていき，$\hat{T}$ 個の時刻 $\hat{t}$ の画像 $\boldsymbol{S}(\hat{t})$ が集まると，一個の短動画 $\hat{\mathcal{S}}$ とする．3.3 で述べた時刻の集合 $\mathcal{T}$ の定義により，インタラクション動画 $\mathcal{S}$ に含まれる時刻の総数は T である．よって，一個の短動画に含まれる時刻の総数を $\hat{T}(< T)$ とする．学習時には，予め準備された L 個のインタラクション動画から生成された LK 個の短動画を用いて C3D ネットワークを訓練する．予測時には，インタラクション動画から生成された K 個の短動画を用いて，会話の活発さを表すラベル候補を K 回算出し，それらの間で最大回数選択されたラベル候補を最終的に出力ラベルとする．

4. 実験

4.1 評価データセット

提案手法の有効性を確認するため，屋外で歩行中の人物グループが会話する様子をカメラを用いて撮影した．撮影環境を図 **5**(a) に示す．路面からカメラまでの高さを 21.4 m とし，屋外駐車場を俯瞰する形で撮影した．図 **5**(a) 右上に示すカメラ動画の視野となるようにカメラ (SONY, FDR-AX55) を取り付けた．カメラの解像度を 3840×2160 画素とし，フレームレートを 30 fps とした．3.5 で述べた路面座標系を図 **5**(b) とした．射影変換を適用するためのホモグラフィ行列を，路面上の白線の交点 4 つから算出した．なお，路面座標系におけるカメラの位置を (10.7, 59.3, 21.4) とした．

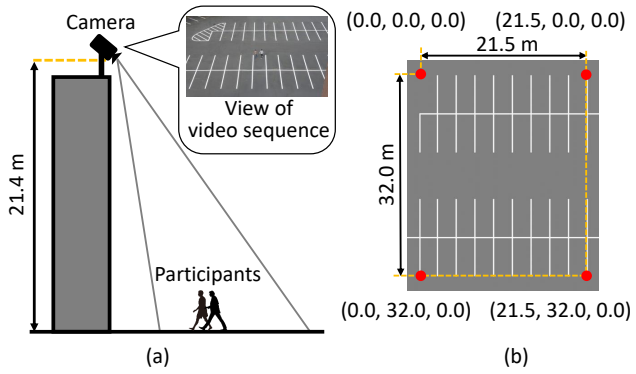


Fig. 5 Setting for collecting video sequences of conversations.

実験協力者を 20 名 (男性 19 名, 女性 1 名, 平均年齢 22.6±1.3 歳) とした．本論文の実験では，人物グループを構成する人数を，会話が発生する最小限の 2 名で統制した．実験協力者 20 名の中から重複なしランダムで 2 名選出し，1 組の人物グループを設定した．最終的に，人物グループを 52 組準備し，カメラ動画を撮影した．なお，実験協力者を集める際，初対面で会話が弾まないことを避けるため，実験協力者同士がある程度の面識があることを条件とした．具体的には同じ大学の同じ学科に所属していることとした．また，それぞれの人物グループでは，2 名が横に並んだ状態で歩行するように統制した．歩行中の人物の配置はいくつかのパターンが想定されるが，ここでは発生頻度が高いと考えられる横に並んだ状態で基本性能を検証した．

それぞれの会話の活発さラベル（活発，非活発，会話なし）について，人物グループが歩行する様子を撮影した．撮影したカメラ動画の例を図 **6** に示す．なお，図中では人物グループの領域のみを拡大表示している．会話の活発さラベルの条件を以下に設定した．

活発

会話のトピックとして，歩行中に互いの趣味について紹介し合うよう指示した．ある人物が趣味について話し，それに対して相手が反応する様子や，その相手が新たに話し始め，元の人物が反応する様子を撮影した．会話を開始する人物はどちらでも良いとした．

非活発

会話のトピックとして，歩行中は互いに興味や関心の薄い話題で会話をするよう指示した．ここでのトピックは，事前に用意した複数のトピック候補 (訪問したことがない国の経済状況や政治情勢など) から，実験協力者が選ぶこととした．会話を開始する人物はどちらでも良いものとした．

会話なし

歩行中に一切会話を行わないよう指示した．

実験協力者には，会話のトピックのみを指示し，身体インタラクションに関して特に説明をせず指示をしなかった．実験協力者には，人物グループを構成する 2 名からなる組の中で，横に並んだ状態で歩行経路を通るよう指示した．それぞれの歩行経路の開始地点を，カメラ動画中の上，下，右上，左下とした．具体的には，図 **5**(b) の路面座標における (10.8, 0.0, 0.0)，(10.8, 32.0, 0.0)，(16.8, 0.0, 0.0)，(0.0, 32.0, 0.0) の位置とした．また，それぞれの歩行経路の目標地点を，(10.8, 32.0, 0.0)，(10.8, 0.0, 0.0)，(0.0, 32.0, 0.0)，(16.8, 0.0, 0.0) の位置とした．各歩行経路で撮影する順番をランダムとした．また，2 名が横に並ぶ順番をランダムとした．最終的に，カメラ動画を 52(組数) × 3(ラベル数) × 4(歩行経路数) = 624 個撮影した．一つのカメラ動画では，実験協力者 2 名の組が，カメラ視野内に出現する開始時刻から消失する終了時刻までとした．

撮影されたカメラ動画の時間長のヒストグラムを図 **7** に示す．カメラ動画の時間長は 12 秒から 30 秒の範囲であった．また，その時間長の平均は 19.7±3.0 秒であった．会話の活発さラベルごとの時間長の平均は，活発の場合が 20.2±3.0 秒，非活発の場合が 20.3±3.0 秒，会話なしの場合が 18.7±2.7 秒であった．なお，3.3 で述べた集合 $\mathcal{T}$ に属する時刻の総数 T の平均は，カメラのフレームレートが 30 fps であるため 591.6 ± 90.2 個であった．

4.2 実験条件

提案手法の評価において，実験時に設定した条件を述べる．まず3.2 で述べた手続き P1 から P3 までの設定について述べる．人物の領域画像を推定する手続き P1 の Mask R-CNN [16] は，姿勢の時系列信号を推定する手続き P2 の PHALP [17] から呼び出されており，その設定を PHALP でのデフォルト値とした．また，手続き P2 における姿勢推定ネットワークモデルも PHALP のデフォルト値とした．人体モデル SMPL の体型パラメータも PHALP のデフォルト値とした．手続き P2 と P3 における Hampel フィルタの窓サイズを 5 とした．手続き P3 の式 (4) の σ を，人物の外接矩形の縦幅に応じて決定した．具体的には，縦幅が大きくなると σ を大きくし，縦幅が小さくなると σ を小さくした．

手続き P4 において，インタラクション動画を生成するための仮想カメラ視点の決め方について述べる．人物グループの真正面の定位置へ常にどの時刻においても，仮想カメラ視点を設置した．このために，一つの人物グループに属する 2 名の人物 p_1, p_2 について，路面上の歩行位置を $\tilde{\boldsymbol{f}}(t, p_1) \in \mathcal{F}(p_1)$, $\tilde{\boldsymbol{f}}(t, p_2) \in \mathcal{F}(p_2)$ とし，路面上での人物グループの中心位置 $(\tilde{\boldsymbol{f}}(t, p_1) + \tilde{\boldsymbol{f}}(t, p_2))/2 = \tilde{\boldsymbol{f}}(t, p_1, p_2)$ を求めた．時刻の集合 $\mathcal{T}$ に属する全ての t における中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ を用いて，直線の当てはめを実行することで，人物グループが路面上で歩行する進行方向を求めた．人物グループの進行方向へ中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ から 4.25 m 離れた位置へ，どの時刻においても仮想カメラ視点を設置した．また，仮想カメラ視点の高さを路面から 0.85 m とした．生成されたインタラクション動画 $\mathcal{S}$ の例を図 **8** に示す．各人物の配色を薄赤か薄青とし，重複無しランダムに与えた．人物グループの正面から光を常に照射するため，光源位置を仮想カメラ視点と連動させた．レンダリングツールとして Blender を用い，そのエンジンとして Workbench を用いた．

手続き P5 における C3D [19] のネットワーク構成を図 **9** に示す．ここでは三次元畳み込み層を 4 個，三次元プーリング層を 4 個，全結合層を 2 個とした．三次元畳み込みのフィルタサイズを $3 \times 3 \times 3$ とした．短動画 $\hat{\mathcal{S}}$ を構成する時刻 $\hat{t}$ の総数 $\hat{T}$ を 16 とした．短動画の配列サイズを 100(画素数) × 100(画素数)× 3(色数) × 16(時刻数) とした．また，3.7 で述べたパラメータを $I = 18$，$K = 50$ とした．C3D の学習時にオプティマイザとして RMSprop を用い，学習率を 0.0001，ミニバッチ

Fig. 6 Examples of camera video sequences representing conversation activities in the groups of pedestrians.

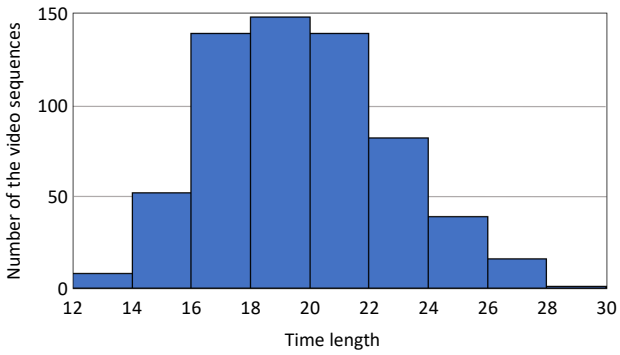


Fig. 7 Histogram of the time length (seconds) of video sequences of conversation activities.

サイズを 16 に設定した．

会話の活発さ認識の精度を求める際に Leave-one-pair-out を適用した．具体的には4.1 で述べた実験協力者 2 名で構成される 1 組から撮影されたインタラクション動画 12 個をテストに用い，残りの 51 組から撮影されたインタラクション動画 $L = 612$ 個を訓練に用いた．この訓練とテストとを全 52 組について繰り返した．なお，1 組あたり 3(ラベル数) × 4(歩行経路数) = 12 個のインタラクション動画を準備した．評価データセットでは，実験協力者 20 名からランダムに選択された 2 名で各組が構成されているため，52 組の中でどちらか片方の人物が，テストと訓練との間で同一になる組が存在した．本論文では，組の中で片方が同じ人物であっても，相手の人物が異なる場合，その組の中で発生する会話の内容や流れが変わると考えた．このため，片方が同一人物であったとしても，各組で生じている身体インタラクションは違っていると仮定し実験を行った．組を構成する両名ともが，テストと訓練との間で同一になる場合は評価データセットには含めなかった．なお，ランダム選択で複数回選ばれた実験協力者には，会話のトピックが同じにならないよう指示した．

4.3 基本性能

インタラクション動画を特徴量として用いる提案手法の有効性を評価した．比較のため，以下の特徴量を用いて会話の活発さ認識の精度を算出した．

M1: **インタラクション動画** $\mathcal{S}$

提案手法のインタラクション動画 $\mathcal{S}$ を特徴量とした．具体的には，手続き P4 で生成された $\mathcal{S}$ から，さらに手続き P5 で短動画 $\hat{\mathcal{S}}$ を生成した．その短動画の配列サイズを 100(画素数) × 100(画素数) × 3(色数) × 16(時刻数) とした．

M2: **人物グループ動画** $\mathcal{R}'$

人物グループの見え方を表す動画 $\mathcal{R}'$ を特徴量とした．その見え方の例は先に図 **6** で示したものである．手続き P1 で推定された領域画像 $\boldsymbol{R}(t,p) \in \mathcal{R}(p)$ を用いて，同一グループに属する 2 名の人物 p_1, p_2 が同一視野内に収まるように，カメラ動画に対して ROI を設定した．同一グループを構成するかどうかを決める際，手続き P4 で述べた人物間の距離を用いた．人物グループ動画 $\mathcal{R}'$ を手続き P5 の認識処理へ直接渡し，さらに人物グループ動画から短動画を生成した．短時系列信号の初期時刻を，M1 と同様にランダムに設定した．その短動画の配列サイズを 100(画素数) × 100(画素数) × 3(色数) × 16(時刻数) とした．

M3: **姿勢の時系列信号** $\mathcal{V}(p)$

人物グループに属する各人物から推定された姿勢の時系列信号 $\mathcal{V}(p)$ を特徴量とした．具体的には，手続き P2 で推定された $\mathcal{V}(p)$ を，手続き P5 の認識処理へ直接渡し，短時系列信号を生成した．短時系列信号の初期時刻を，同様にランダムに設定した．その短時系列信号の配列サイズを 6890(頂点数) × 2(人数) × 3(成分数) × 16(時刻数) とした．なお，他の特徴量と次元数を揃えるために，一様乱数に従い頂点数を 5000 にランダムサンプリングした後，配列サイズを $5000 \times 2 \times 3 \times 16$ から $100 \times 100 \times 3 \times 16$ へ

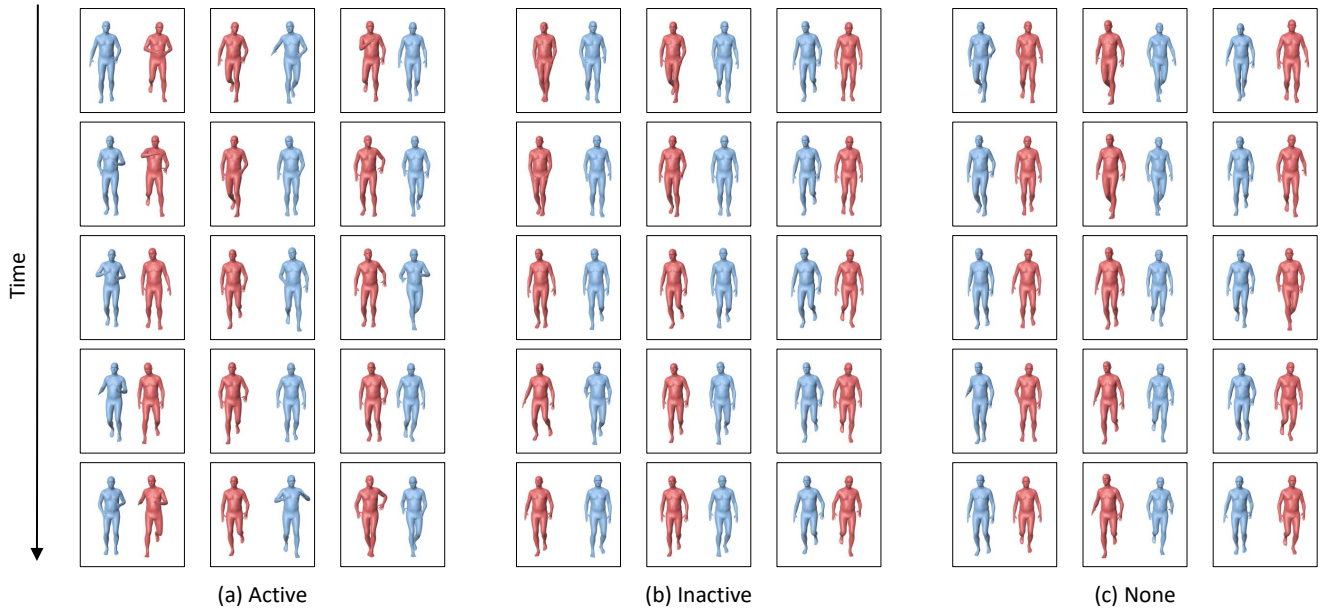


Fig. 8 Examples of interaction video sequences $\mathcal{S}$

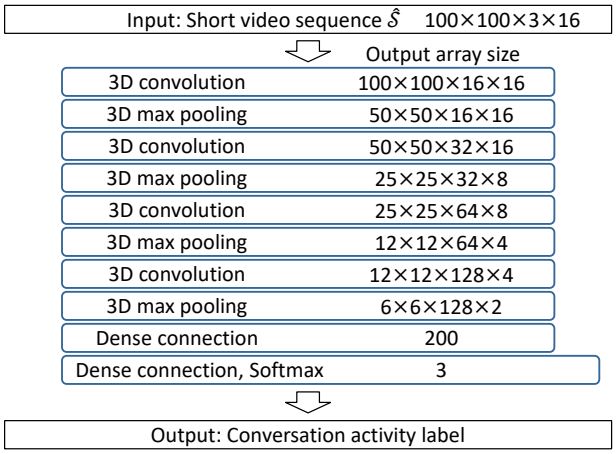


Fig. 9 Architecture of C3D in our experiments

変形した．精度評価の各試行において，全ての時系列信号でサンプリングされる頂点は同じとした．

M4: **姿勢と歩行位置からなる時系列信号** $\tilde{\mathcal{V}}(p)$

人物グループに属する各人物から推定された姿勢の時系列信号 $\mathcal{V}(p)$ と歩行位置の時系列信号 $\mathcal{F}(p)$ とを統合したものを特徴量とした．具体的には，手続き P4 で生成されたレンダリング前の時系列信号 $\tilde{\mathcal{V}}(p)$ を，手続き P5 の認識処理へ直接渡し，短時系列信号を生成した．短時系列信号の初期時刻を，同様にランダムに設定した．その短時系列信号の配列サイズを 6890(頂点数) × 2(人数) × 3(成分数) × 16(時刻数) とした．なお，他の特徴量と次元数を揃えるために，一様乱数に従い頂点数を 5000 にランダムサンプリングした後，配列サイズを $5000 \times 2 \times 3 \times 16$ から $100 \times 100 \times 3 \times 16$ へ変形した．精度評価の各試行において，全ての時系列信号でサンプリングされる頂点は同じとした．

以下では，会話の活発さラベルが正しく認識できたかどうかの精度を評価指標とした．手続き P5 のクラス分類ネットワーク

Table 1 Comparison of accuracy (%) using each feature for conversation activity recognition

Feature for conversation activity recognition	Accuracy
M1: Interaction video sequence $\mathcal{S}$	**76.2±0.7**
M2: Pedestrian group video sequence $\mathcal{R}'$	57.3±1.3
M3: Temporal pose signal $\mathcal{V}(p)$	72.9±0.9
M4: Temporal pose and position signal $\tilde{\mathcal{V}}(p)$	74.1±0.7

である C3D へ，それぞれの特徴量を入力し，会話の活発さ認識の精度を算出した．各特徴量を抽出する際にランダムサンプリングがあるため，認識精度を算出する試行回数を 10 回とした．その他の実験条件を4.2 で述べたものとした．

会話の活発さ認識において，各特徴量を用いた場合の精度の比較を表 **1** に示す．M1 のインタラクション動画 $\mathcal{S}$ では認識精度が 76.2±0.7% であり，M2 の人物グループ動画 $\mathcal{R}'$ では 57.3±1.3%, M3 の姿勢の時系列信号 $\mathcal{V}(p)$ では 72.9±0.9%, M4 の姿勢と歩行位置からなる時系列信号 $\tilde{\mathcal{V}}(p)$ では 74.1±0.7% であった．いずれの場合においても，提案手法のインタラクション動画を用いた時の精度が高いことが確認できた．以上の結果より，提案手法で生成されたインタラクション動画を特徴量として用いることは，人物グループ動画や姿勢と歩行位置の時系列信号を特徴量として用いることに比べて，会話の活発さ認識に有効であると言える．

4.4 時系列信号を取り扱う深層学習の手法を適用した場合の精度評価

会話の活発さ認識において，時系列信号を取り扱う深層学習の手法を適用した場合について評価した．手続き P5 において，前節の実験では動画を取り扱うために開発された C3D を用いたが，本節では時系列信号を取り扱うために開発された LSTM [20] や GRU [21] を用いた．LSTM や GRU へ入力する特徴量を以下とした．

M3': **姿勢パラメータの時系列信号** $\mathcal{V}'(p)$

人物グループに属する 2 名の人物について，手続き P2 の

PHALP の内部で推定された姿勢パラメータの時系列信号そのものを特徴量とした．具体的には，人体の 23 個の関節点における回転行列パラメータの時系列信号 $\mathcal{V}'(p)$ を，手続き P5 の認識処理へ直接渡し，短時系列信号を生成した．短時系列信号の初期時刻をランダムに設定し，時刻の刻み幅 $I = 18$ として，短時系列信号を $K = 50$ 個生成した．その信号の 2 次元配列サイズを (23(関節点数) · 3(回転成分数) · 2(人数)) × 16(時刻数) とした．

M3”: **歩行位置の時系列信号** $\mathcal{F}(p)$

人物グループに属する 2 名の人物について，手続き P3 で推定された路面上の歩行位置の時系列信号そのものを特徴量とした．具体的には，歩行位置の時系列信号 $\mathcal{F}(p)$ を，手続き P5 の認識処理へ直接渡し，短時系列信号を生成した．その生成手続きを M3’ と同じとした．その信号の 2 次元配列サイズを (2(位置成分数) · 2(人数)) × 16(時刻数) とした．

M4’: **姿勢パラメータと歩行位置の時系列信号** $\mathcal{V}'(p) + \mathcal{F}(p)$

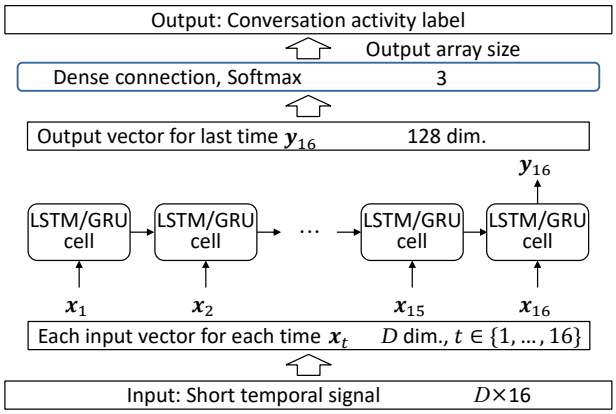
人物グループに属する 2 名の人物について，手続き P2 の PHALP の内部で推定された姿勢パラメータの時系列信号 $\mathcal{V}'(p)$ に，歩行位置の時系列信号 $\mathcal{F}(p)$ を結合したものを特徴量とした．この結合のため，それぞれの時系列信号を表す特徴ベクトルを単純に連結し，手続き P5 の認識処理へ直接渡すことで，短時系列信号を生成した．その生成手続きを M3’ と同じとした．その信号の 2 次元配列サイズを ((23(関節点数) · 3(回転成分数) + 2(位置成分数)) · 2(人数)) × 16(時刻数) とした．

それぞれの特徴量を LSTM または GRU へ入力し，会話の活発さ認識精度を算出した．認識精度を算出する試行回数を各特徴量で 10 回とした．LSTM，または，GRU を会話の活発さ認識に用いる場合のネットワーク構成を図 **10** に示す．図中の $\boldsymbol{x}_t$ は，入力された短時系列信号において，時刻 t の要素のみで表されるベクトルとした．なお $\boldsymbol{x}_t$ の次元数 D を，上記の特徴量 M3’, M3”, M4’ の説明で述べた配列サイズを用いて，それぞれ設定した．次元数 D の具体的な値を図 **10** 下部に示す．特徴量の間で時刻数を 16 に統一した．LSTM を認識に用いる場合では LSTM セルを時系列に 16 個連結し，GRU を認識に用いる場合では GRU セルを時系列に 16 個連結した．最終時刻の出力ベクトル $\boldsymbol{y}_{16}$ の次元数を 128 とし，そのベクトルを全結合層に入力することで，活発さラベルを予測した．いずれも学習時にオプティマイザとして RMSprop を用いた．LSTM での学習率を，特徴量 M3’ で 0.00001，特徴量 M3” で 0.005，特徴量 M4’ で 0.00001 とした．GRU での学習率を，特徴量 M3’ で 0.00001，特徴量 M3” で 0.005，特徴量 M4’ で 0.001 とした．ミニバッチサイズを 32 とした．その他の実験条件を4.2 で述べたものとした．

会話の活発さ認識において，時系列信号を取り扱う深層学習の手法を適用した場合の精度を表 **2** に示す．提案手法のインタラクション動画を用いた特徴量 M1 と C3D とを組み合わせた場合は，姿勢パラメータや歩行位置の時系列信号を用いた特徴量 M3’, M3”, M4’ と LSTM や GRU とを組み合わせた場合と比べて，認識精度が高いことが確認できた．

4.5 仮想カメラ視点を変えた場合の評価

インタラクション動画を手続き P4 において生成する時の仮想カメラ視点を変えた場合について，会話の活発さ認識の精度を評価した．比較に用いる仮想カメラ視点の位置を以下とした．



D=23･3･2 in M3’, 2･2 in M3”, (23･3+2)･2 in M4’

Fig. 10 Architecture when using LSTM or GRU in our experiments

Table 2 Accuracy (%) of conversation activity recognition when applying deep learning methods for handling time series signals

Feature	Method	Accuracy
M1: Ours	C3D	**76.2±0.7**
M3’: Pose parameter $\mathcal{V}'(p)$	LSTM	67.4±0.9
M3’: Pose parameter $\mathcal{V}'(p)$	GRU	67.6±1.0
M3”: Foot position $\mathcal{F}(p)$	LSTM	40.8±1.0
M3”: Foot position $\mathcal{F}(p)$	GRU	40.5±1.2
M4’: $\mathcal{V}'(p) + \mathcal{F}(p)$	LSTM	67.1±1.1
M4’: $\mathcal{V}'(p) + \mathcal{F}(p)$	GRU	67.5±0.9

C1: **正面**

人物グループの正面へ仮想カメラを常に固定した．ここでは 4.2 で述べた人物グループの中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ を用いた．歩行している進行方向において，中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ から 4.25 m 離れた路面上の位置へ，どの時刻においても仮想カメラ視点を設置した．その視点の高さを路面から 0.85 m とした．この仮想カメラ視点の設置条件は 4.2 で述べた提案手法そのものである．

C2: **背面**

人物グループの背面へ仮想カメラを常に固定した．歩行している進行方向において，その人物グループの中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ から -4.25 m 離れた路面上の位置へ，どの時刻においても仮想カメラ視点を設置した．その視点の高さを路面から 0.85 m とした．

C3: **頭上**

人物グループの頭上へ仮想カメラを常に固定した．人物グループの路面上での中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ へ，どの時刻においても仮想カメラ視点を設置した．その視点の高さを路面から $0.85 + 4.25 = 5.1$ m とした．

C4: **足下**

人物グループの足下へ仮想カメラを常に固定した．人物グループの路面上での中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ へ，どの時刻においても仮想カメラ視点を設置した．その視点の高さを路面から $0.85 - 4.25 = -3.4$ m とした．

C5: **右側面**

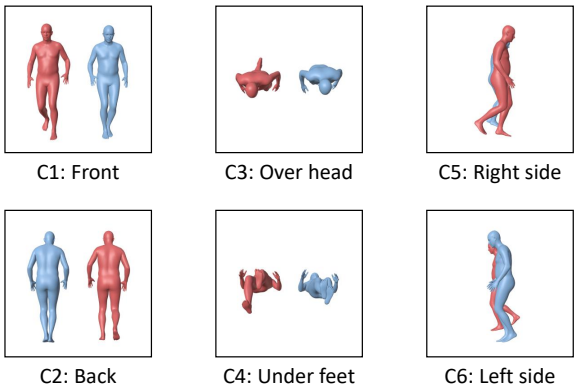


Fig. 11 Examples of interaction video sequences generated from different virtual camera viewpoints in P4

Table 3 Accuracy (%) of conversation activity recognition when generating interaction video sequences from different virtual camera viewpoints in P4

Virtual camer viewpoint	Accuracy
C1: Front	**76.2±0.7**
C2: Back	74.8±0.3
C3: Over head	70.2±0.8
C4: Under feet	72.0±0.9
C5: Right side	40.0±1.5
C6: Left side	48.2±2.6

人物グループの右側面へ仮想カメラを常に固定した．歩行している進行方向と直交する方向へ，その人物グループの中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ から 4.25 m 離れた路面上の位置へ，どの時刻においても仮想カメラ視点を設置した．その視点の高さを路面から 0.85 m とした．

C6: 左側面

人物グループの左側面へ仮想カメラを常に固定した．歩行している進行方向と直交する方向へ，その人物グループの中心位置 $\tilde{\boldsymbol{f}}(t, p_1, p_2)$ から -4.25 m 離れた路面上の位置へ，どの時刻においても仮想カメラ視点を設置した．その視点の高さを路面から 0.85 m とした．

それぞれの仮想カメラ視点を用いて生成されたインタラクション動画の例を図 **11** に示す．本節の実験では，仮想カメラ視点の位置のみを変更し，その他の実験条件を4.2 で述べたものとした．認識精度を算出する試行回数を各仮想カメラ視点で 10 回とした．

インタラクション動画を生成する時に仮想カメラ視点を変えた場合において，会話の活発さ認識の精度を評価した結果を表 **3** に示す．仮想カメラ視点を人物グループの正面とした場合 C1 は，正面以外の場合 C2, C3, C4, C5, C6 と比べて，認識精度が高いことを確認した．以上の結果より，インタラクション動画を手続き P4 で生成する時，人物グループを常に正面から捉える位置に仮想カメラ視点を設置することは，会話の活発さ認識に有効であると言える．

4.6 提案手法における短動画生成のパラメータ評価

提案手法の手続き P5 において，短動画を生成する際のパラメータについて，会話の活発さ認識の精度を評価した．ここでは3.7 で述べたパラメータ I および K を評価の対象とした．

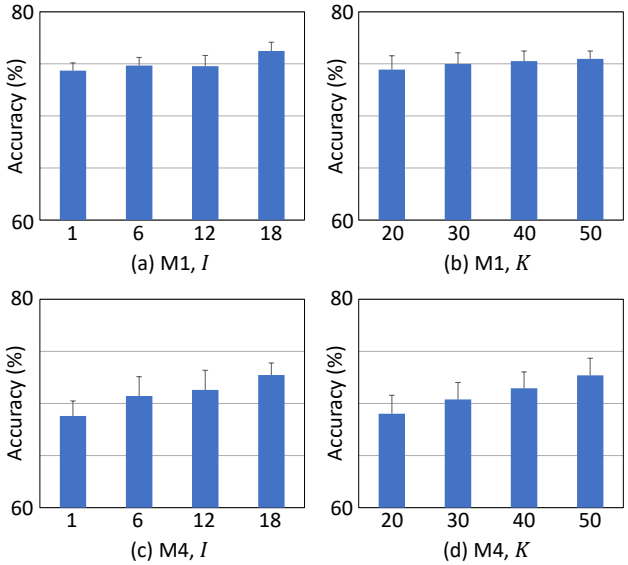


Fig. 12 Evaluation of the accuracy when changing the parameters for generating short video sequences in P5

$I = 1, 6, 12, 18$，$K = 20, 30, 40, 50$ とした時の組み合わせ全 16 通りについて認識精度を算出した．その試行回数を，各組み合わせで 5 回とした．表 **1** にて，最も精度が高い提案手法の特徴量 M1 と，次に精度が高い比較手法の特徴量 M4 とについて，上記パラメータを評価した．その他の実験条件を4.2 で述べたものとした．

まず，提案手法の特徴量 M1 における短動画生成のパラメータ I を変えた時の活発さ認識の精度を図 **12**(a) に示す．図中のグラフでは，I のあるパラメータ値に対して $K = 20, 30, 40, 50$ とした時の平均精度を用いた．棒グラフに付属するバーは標準偏差を表す．パラメータ I が大きくなるにつれて認識精度が徐々に向上し，$I = 18$ で最も高くなった．次に，提案手法における短動画生成のパラメータ K を変えた時の活発さ認識の精度を図 **12**(b) に示す．図中のグラフでは，K のあるパラメータ値に対して $I = 1, 6, 12, 18$ とした時の平均精度を用いた．棒グラフに付属するバーは標準偏差を表す．$K = 20$ と比較して，$K = 30, 40, 50$ で認識精度が僅かではあるが向上した．パラメータの組み合わせとして $I = 18$ および $K = 50$ で，会話の活発さ認識の精度が最も高くなっていた．

短動画生成のパラメータの評価結果について考察する．パラメータ I が大きくなるにつれて認識精度が徐々に向上したことから，短動画を生成する際，元のインタラクション動画から，時間方向の間隔を広くとり画像をサンプリングした方がよいと言える．このことから，$I = 1$ の場合など，時間方向に速く変化する高周波の身体インタラクションよりも，$I = 18$ の場合など，遅く変化する低周波の身体インタラクションが，会話の活発さ認識に有効であると考えられる．この遅く変化する身体インタラクションは，グループ内の会話の流れに応じて徐々に変化していく人物らの動きを捉えていると考えられる．また，パラメータ K を大きくすることで僅かに認識精度が向上する傾向が見られたことから，複数の短動画の間で一部の画像が重複しても構わないので，より多くの個数の短動画を生成した方がよいと考えられる．これは，同一の活発さラベルの中で，短動画の初期時刻のずれを，訓練時になるべく学習した方がよいためと考えられる．以上より，会話の活発さを精度よく認識で

きる特徴量として，時間方向に遅く変化する身体インタラクションについて，初期時刻を多様にずらしつつ短動画を生成すればよいと言える．

次に，比較手法の特徴量 M4 における短動画生成のパラメータ I を変えた時の活発さ認識の精度を図 **12**(c) に示す．図中のグラフでは，I のあるパラメータ値に対して $K = 20, 30, 40, 50$ とした時の平均精度を用いた．棒グラフに付属するバーは標準偏差を表す．パラメータ I が大きくなるにつれて認識精度が徐々に向上し，$I = 18$ で最も高くなった．次に，提案手法における短動画生成のパラメータ K を変えた時の活発さ認識の精度を図 **12**(d) に示す．図中のグラフでは，K のあるパラメータ値に対して $I = 1, 6, 12, 18$ とした時の平均精度を用いた．棒グラフに付属するバーは標準偏差を表す．パラメータ K が大きくなるにつれて認識精度が徐々に向上し，$K = 50$ で最も高くなった．パラメータの組み合わせとして $I = 18$ および $K = 50$ で，会話の活発さ認識の精度が最も高くなっていた．

上記の実験結果から，比較手法の特徴量 M4 と提案手法の特徴量 M1 のいずれにおいても，パラメータ I および K が大きくなるにつれて，会話の活発さ認識の精度が向上するという傾向が同様に見られた．また，どのパラメータの組み合わせにおいても，M1 の認識精度が，M4 の認識精度より総じて高かった．M1 と M4 とは，4.3 で述べた通り，いずれも同一の時系列信号から生成された特徴量である．両者で異なる点は，それらの信号からインタラクション動画を生成するか，そのまま時系列信号を適用するかである．このことから，インタラクション動画を特徴量として用いることが，パラメータの組み合わせを変えたとしても，会話の活発さ認識に有効であることが示唆される．

4.7 提案手法の実験結果に関する解釈と考察

ここまでの実験結果より，カメラ動画から抽出されたインタラクション動画を特徴量とする提案手法の精度が高いことが確認された．ここでは実験結果の解釈と考察とを行うため，4.3 で述べた提案手法の特徴量 M1 を用いた場合において混同行列を求めた．この混同行列の全要素の合計数を 52(組数) × 3(ラベル数) × 4(歩行経路数) × 10(試行数) = 6240 とした．提案手法の特徴量 M1 の認識精度は表 **1** で示した通り 76.2±0.7% であった．その混同行列を表 **4** に示す．まず，対角線上の正解数を見ると，活発が最も多く，次いで会話なしと非活発であった．ただし非活発と会話なしとの間で正解数は近かった．活発では，どの人物グループでも身体インタラクションが多く生じており，非活発や会話なしと比べて認識しやすいと考えられる．次に誤り数を見ると，非活発は会話なしに間違えやすく，会話なしは非活発に間違えやすいことが分かった．非活発では，身体インタラクションがほぼ生じていない人物グループが一定数存在したため，会話なしと近い特徴分布になったと考えられる．実際にカメラ動画を視認すると，非活発では会話を途中で諦める人物と継続しようとする人物がおり，グループに属する人物の個性に応じて，身体インタラクションが変わることの影響が大きいと考えられる．なお活発でも，非活発や会話への間違いがある程度発生していた．これも人物の個性に応じた身体インタラクションの変化が影響していると考えられる．実際に活発ラベルではあるが，それほど大きなジェスチャをとらない人物や，相手の方に向かずに進行方向を見ている人物が存在した．また，非活発ラベルであるが，大きなジェスチャをとる人物や，相手の方へ向く頻度が高い人物が存在した．会話における身体インタラクションには個性の影響が大きいと考えられるため，今後の課題として，個性に頑健な認識手法の開発が考えられる．また，本論文ではグループ単位で会話の活発さを認識したが，グループに属する人物ごとに会話の活発さを認識することも考えられる．

Table 4 Confusion matrix when using the feature M1 of our method

		Predicted label		
		Active	Inactive	None
True label	Active	1949	90	41
	Inactive	102	1381	597
	None	31	623	1426

5. 会話の活発さ認識における本論文の価値と今後の課題

本論文では，屋外で歩行中の人物グループで発生する会話の活発さを，カメラ動画から獲得される身体インタラクションを用いて認識する手法を設計した．会話の活発さ認識における本論文の価値を考えるため，会話中に生じる身体インタラクションにおける変動要因を以下に挙げる．各要因において括弧内に評価データセットを構築した時のパラメータを示す．

要因 1　人物グループを構成する人数 (2 名)
要因 2　歩行中の人物の並び方 (横並び)
要因 3　グループ内の人物間の関係性 (互いに面識がある関係)
要因 4　各人物が置かれている状況 (普段の日常)

データセット構築時には発生頻度が多いと考えられるパラメータを設定した．これらのパラメータで最小単位の歩行者の組を構成し，4.1 で述べた会話のトピックを用いて活発さラベルを決め，カメラ動画を撮影した．この独自に収集したデータセットを用いることで，屋外で歩行中の人物グループにおける会話の活発さについて，認識精度を初めて評価できた価値はあると考える．

ただし上記の変動要因は，実環境では多様に変化するため，さらに考慮すべきことが多く存在する．以下では，屋外で歩行中の人物グループで発生する会話の活発さを認識する際，それぞれの変動要因で想定されることを列挙する．

要因 1　3 名以上のグループも存在する．さらに，途中からの合流や離脱で人数が増減することがある．グループを構成する人数が多い場合は，サブグループに分かれた身体インタラクションが発生することがある．
要因 2　縦並びや斜め並びも存在する．その並び方で歩行する上で，本論文で用いた直進だけでなく湾曲など経路に多様性があり，グループ間のすれ違いが発生することがある．また，グループ内の人物間で相対的な移動速度が異なると，並び方が時々刻々と変化していくこともある．
要因 3　家族，恋人，友人など様々な関係がある．さらに国際的な文化圏の違いなども考えていく必要がある．
要因 4　極端な状況に置かれることで，ストレスや負の感情や喜びの感情などが高まり身体インタラクションが変わると考えられる．

上記の変動要因のどれを優先的に取り組むかを，応用に合わせて熟慮しつつ，歩行中の人物グループ内で生じる身体インタラ

クションとは何であるかを追求し，新たな評価データセットを構築していく必要があると考える．

さらに，人物グループが存在している周囲環境に起因する変動要因を3.1 で述べたように考慮する必要がある．屋外で歩行する人物グループを対象としているため，時間帯や天候や季節により照明条件が異なり，人物グループ内で生じる身体インタラクションの見え方が劇的に変化していく．カメラを取り付けることができる設置場所の制約で俯角や画角が異なり，身体インタラクションの見え方が変化していく．また，カメラのフレームレートが遅い場合は，有効な時系列信号を推定できない可能性が考えられる．さらに，自動車や自転車の出入りや，同時出現人数が多くなることで，一部の人物が隠れてしまう場合が発生する．これらの変動要因は，本論文で扱った会話の活発さ認識だけでなく，様々な画像認識の屋外応用で共通である．実環境に耐え得る技術を目指し，評価データセットを拡充しつつ，それぞれの変動要因に取り組んでいく必要がある．

6. おわりに

本論文では，屋外で歩行中の人物グループを対象とし，カメラ動画から抽出されたインタラクション動画を用いて，会話の活発さを認識する手法について述べた．提案手法では，人物グループに属する各人物の姿勢と位置との時系列信号を用いて，その人物グループ内で生じている身体インタラクションを表す動画を仮想空間で生成した．このインタラクション動画を特徴量とすることで，会話の活発さラベルを動画分類のニューラルネットワークを用いて決定した．実験結果より，提案手法のインタラクション動画を特徴量として用いることで，人物領域の動画を特徴量として用いる場合や，姿勢の時系列信号と歩行位置の時系列信号とを特徴量として用いる場合に比べて，会話の活発さを精度良く認識できることを確認した．

今後の課題として，会話の活発さをより詳細に認識できる手法の開発や，隠れに頑健な歩行位置の推定手法の開発が挙げられる．また，人物グループに属する人数が増えた場合の評価，人物グループ内での人物の配置が変化した場合の評価，カメラ動画中に複数の歩行者グループが存在する場合の評価が挙げられる．

謝辞

本研究を進めるにあたり，貴重なご意見やご指摘をくださいました西菱電機株式会社の鳥居紀彦様，三宅智博様，吉村修様，尾嵜憲司様，鳥取大学工学部教授の岩井儀雄先生に感謝の意を表する．

参 考 文 献

1) D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.
2) F. Zanlungo, D. Brščić, and T. Kanda. Pedestrian group behaviour analysis under different density conditions. *Transportation Research Procedia*, Vol. 2, pp. 149–158, 2014.
3) I. Chamveha, Y. Sugano, Y. Sato, and A. Sugimoto. Social group discovery from surveillance videos: A data-driven approach with attention-based cues. In *Proceedings of the British Machine Vision Conference*, pp. 1–12, 2013.
4) W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 5, pp. 1003–1016, 2012.
5) M. Zanotto, L. Bazzani, M. Cristani, and V. Murino. Online bayesian nonparametrics for group detection. In *Proceedings of the British Machine Vision Conference*, pp. 1–12, 2012.
6) F. Solera, S. Calderara, and R. Cucchiara. Socially constrained structural learning for groups detection in crowd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 5, pp. 995–1008, 2016.
7) J. Su, J. Huang, L. Qing, X. He, and H. Chen. A new approach for social group detection based on spatio-temporal interpersonal distance measurement. *Heliyon*, Vol. 8, No. 10, p. e11038, 2022.
8) T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems*, Vol. 1, p. 1216–1224, 2010.
9) P. Rota, N. Conci, and N. Sebe. Real time detection of social interactions in surveillance video. In *Proceedings of the European Conference on Computer Vision*, pp. 111–120, 2012.
10) S. Odashima, M. Shimosaka, T. Kaneko, R. Fukui, and T. Sato. Collective activity localization with contextual spatial pyramid. In *Proceedings of the European Conference on Computer Vision*, pp. 243–252, 2012.
11) R. Han, H. Yan, J. Li, S. Wang, W. Feng, and S. Wang. Panoramic human activity recognition. In *Proceedings of the European Conference on Computer Vision*, pp. 224–261, 2022.
12) W. Kraaij, T. Hain, M. Lincoln, and W. Post. The AMI meeting corpus. In *Proceedings of the International Conference on Methods and Techniques in Behavioral Research*, pp. 137–140, 2005.
13) L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang. VACE multimodal meeting corpus. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*, pp. 40–51, 2006.
14) H. Hung and G. Chittaranjan. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the ACM International Conference on Multimedia*, pp. 879–882, 2010.
15) E. Kurtić, B. Wells, G. J. Brown, T. Kempton, and A. Aker. A corpus of spontaneous multi-party conversation in bosnian serbo-croatian and british english. In *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1323–1327, 2012.
16) K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.
17) J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik. Tracking people by predicting 3D appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2740–2749, 2022.
18) M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, Vol. 34, No. 6, pp. 1–16, 2015.
19) D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.
20) S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.
21) J. Chung, C. Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.