追加ラベルを組み込んだ深層生成モデルを用いた時系列骨格情報による人物識別

○浅見直人 （中央大学） Alessandro Moro (RITECS Inc.)
池勇勳 （JAIST） 梅田和昇 （中央大学）

1. はじめに

動画像内に映った歩行者をその歩き方から識別する歩容認証が犯罪捜査などで利用されている．従来の歩容認証手法の多くはシルエット画像に基づくものである [1] [2] が，それらは荷物の所持の違いや服装の違いによる認証精度低下の問題がある．そこで本研究では，動画像から骨格データを推定し，得られた時系列骨格データから人物を識別する．

一般に歩容特徴は，観測方向や歩行速度，歩行周期のばらつきなどの共変量を受け変化するため，共変量による変化を吸収できる認証手法が求められている．また，年齢や性別といった人物ラベル以外のラベルの分類をサブタスクとして加えることで，歩容認証の認証制度を向上させた研究も存在する [3]．しかしその手法では CNN を用いた識別モデルによってクラス分類を行っており，各ラベル間の関係，優先度を考慮できない．そのような共変量や人物ラベル以外のラベルは歩容特徴を形作る潜在変数と考えることができる．本研究では，そのような人物ラベル以外の潜在変数を加えた深層生成モデルによる識別を目指し，人物ラベルと方向ラベルの二つを潜在変数とした生成モデルを用いた変分オートエンコーダ (VAE: Variational autoenoder) を作成する．さらに，各ラベル間の関係を考慮した生成モデルの構築とその VAE によるクラス分類を提案する．

2. 生成モデル

本研究の生成モデルは $\boldsymbol{x}$ を時系列骨格データ，$\boldsymbol{s}_1$ を人物ラベルを表すカテゴリ潜在変数，$\boldsymbol{s}_2$ を方向ラベルを表すカテゴリ潜在変数として，

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1) &= p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1) \\ &= p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1)p(\boldsymbol{\mu}_2|\boldsymbol{s}_2)p(\boldsymbol{s}_1)p(\boldsymbol{s}_2) \end{aligned}$$

とする．この生成モデルでは，方向ラベル $\boldsymbol{s}_2$ が得られた条件付き分布 $p(\boldsymbol{\mu}_2|\boldsymbol{s}_2)p(\boldsymbol{s}_2)$ に，時系列骨格データの特徴のうち方向成分を意味する潜在変数 $\boldsymbol{\mu}_2$ が従い，方向成分 $\boldsymbol{\mu}_2$ と人物ラベル $\boldsymbol{s}_1$ が得られた条件付き分布 $p(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1)p(\boldsymbol{s}_1)$ に，時系列骨格情報全体を表す潜在変数 $\boldsymbol{z}$ が従うような階層的なモデルを仮定する．つまりこの生成モデルでは，時系列骨格データが人物成分より大まかな特徴である方向成分をベースにし人物成分が加わっていると仮定している．ここで，$p(\boldsymbol{\mu}_2|\boldsymbol{s}_2)p(\boldsymbol{s}_2)$，$p(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1)p(\boldsymbol{s}_1)$ はそれぞれ以下の混合ガウス分布である．

$$p(\boldsymbol{\mu}_2|\boldsymbol{s}_2)p(\boldsymbol{s}_2) = \prod_{i=1}^{M}\mathcal{N}(\boldsymbol{\mu}_2|\boldsymbol{m}_i,\Sigma_i)^{\boldsymbol{s}_{2,i}}\boldsymbol{\pi}_2^{\boldsymbol{s}_{2,i}} \tag{1}$$

$$p(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1)p(\boldsymbol{s}_1) = \prod_{j=1}^{K}\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_2+\boldsymbol{\mu}_{1,j},\Sigma_j)^{\boldsymbol{s}_{1,j}}\boldsymbol{\pi}_1^{\boldsymbol{s}_{1,j}} \tag{2}$$

つまり，$\boldsymbol{z}$ が従う混合ガウス分布の平均ベクトルが方向ラベル成分 $\boldsymbol{\mu}_2$ と人物ラベル成分 $\boldsymbol{\mu}_1$ を持つことになる．

本研究で用いる VAE も，一般的に用いられる VAE と同様に事前分布 $p(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1)$ と同じ形の近似事後分布 $q_\phi(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1|\boldsymbol{x})$ を仮定する．

2.1 サンプリング

本研究で用いるネットワークの構造を図 1 に示す．エンコーダは二つのネットワークに分かれる．Encoder2 は時系列骨格データから方向ラベル成分を抽出するためのエンコーダであり，$q_\phi(\boldsymbol{\mu}_2|\boldsymbol{s}_2,\boldsymbol{x})q_\phi(\boldsymbol{s}_2|\boldsymbol{x})$ のパラメータである $\boldsymbol{m}_i,\Sigma_{2,i},\boldsymbol{\pi}_2(i=1,\ldots,M)$ を出力する．この分布は式 (1) で表される混合ガウス分布である．カテゴリカル分布部分 $q_\phi(\boldsymbol{s}_2|\boldsymbol{x})$ は Concrete distribution と呼ばれる連続化したカテゴリカル分布を用い，Gumbel-softmax trck によってサンプリングできる [4]．また，条件付き分布部分 $q_\phi(\boldsymbol{\mu}_2|\boldsymbol{s}_2,\boldsymbol{x})$ は，一つのガウス分布 $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_{\mathrm{gmm2}},\Sigma_{\mathrm{gmm2}})$ に変形でき，その共分散行列 Σ_{gmm2} は

$$\Sigma_{\mathrm{gmm2}} = \left\{\sum_{i=1}^{M} s_{2,i}\Sigma_{2,i}^{-1}\right\}^{-1}$$

で表され，その平均 $\boldsymbol{\mu}_{\mathrm{gmm2}}$ は

$$\boldsymbol{\mu}_{\mathrm{gmm2}} = \Sigma_{\mathrm{gmm2}}\left\{\sum_{i=1}^{M} s_{2,i}\Sigma_{2,i}^{-1}\boldsymbol{m}_i\right\}$$

で表される．結局，$\boldsymbol{\mu}_2$ はガウス分布の reparameterization trick[5] によってサンプリングできる．一方，Encoder1 は得られた方向ラベル成分 $\boldsymbol{\mu}_2$ を利用して時系列骨格データ全体の特徴を抽出するためのエンコーダであり，$q_\phi(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1,\boldsymbol{x})q_\phi(\boldsymbol{s}_1|\boldsymbol{x})$ のパラメータである $\boldsymbol{\mu}_{1,j},\Sigma_{1,j},\boldsymbol{\pi}_1(j=1,\ldots,M)$ を出力する．この分布も式 (2) に示す通り混合ガウス分布であるので，$\boldsymbol{s}_1$ は Gumbel-sofrmax trick によってサンプリングし，$\boldsymbol{z}$ は適切な平均 $\boldsymbol{\mu}_{gmm1}$ と共分散行列 Σ_{gmm1} を用いたガウス分布の reparameterization trick でサンプリングを行う．

2.2 変分下限

本研究で用いる VAE の変分下限は式 (3) で与えられる．

$$\mathcal{L}[\phi,\theta] = \mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}\left[q_\phi(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1|\boldsymbol{x})||p(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1)\right] \tag{3}$$

一般的な VAE と同様に式 (3) の第一項は再構成誤差項であるため，第二項の KL ダイバージェンス項について考える．KL ダイバージェンスは式 (4) のように変形できる．

$$\begin{aligned} &D_{KL}\left[q_\phi(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1|\boldsymbol{x})||p(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1)\right] \\ &= \mathbb{E}_{q_\phi(\boldsymbol{z},\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1|\boldsymbol{x})}\left[\log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1,\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1)}\right] \\ &+ \mathbb{E}_{q_\phi(\boldsymbol{s}_2,\boldsymbol{\mu}_2|\boldsymbol{x})}\left[\log\frac{q_\phi(\boldsymbol{\mu}_2|\boldsymbol{s}_2,\boldsymbol{x})}{p(\boldsymbol{\mu}_2|\boldsymbol{s}_2)}\right] \\ &+ \mathbb{E}_{q_\phi(\boldsymbol{s}_1|\boldsymbol{x})}\left[\log\frac{q_\phi(\boldsymbol{s}_1|\boldsymbol{x})}{p(\boldsymbol{s}_1)}\right] + \mathbb{E}_{q_\phi(\boldsymbol{s}_2|\boldsymbol{x})}\left[\log\frac{q_\phi(\boldsymbol{s}_2|\boldsymbol{x})}{p(\boldsymbol{s}_2)}\right] \end{aligned} \tag{4}$$

ここで，式 (4) の第一項は

$$\begin{aligned} &\mathbb{E}_{q_\phi(\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1|\boldsymbol{x})}\left[\int q_\phi(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1,\boldsymbol{x})\log\frac{q_\phi(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1,\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1)}d\boldsymbol{z}\right] \\ &= \mathbb{E}_{q_\phi(\boldsymbol{s}_2,\boldsymbol{\mu}_2,\boldsymbol{s}_1|\boldsymbol{x})}\left[D_{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1,\boldsymbol{x})||p(\boldsymbol{z}|\boldsymbol{\mu}_2,\boldsymbol{s}_1)\right]\right] \end{aligned}$$

となり，式 (4) の第二項は

$$\begin{aligned} &\mathbb{E}_{q_\phi(\boldsymbol{s}_2|\boldsymbol{x})}\left[\int q_\phi(\boldsymbol{\mu}_2|\boldsymbol{s}_2,\boldsymbol{x})\log\frac{q_\phi(\boldsymbol{\mu}_2|\boldsymbol{s}_2,\boldsymbol{x})}{p(\boldsymbol{\mu}_2|\boldsymbol{s}_2)}d\boldsymbol{\mu}_2\right] \\ &= \mathbb{E}_{q_\phi(\boldsymbol{s}_2|\boldsymbol{x})}\left[D_{KL}\left[q_\phi(\boldsymbol{\mu}_2|\boldsymbol{s}_2,\boldsymbol{x})||p(\boldsymbol{\mu}_2|\boldsymbol{s}_2)\right]\right] \end{aligned}$$

となるため，それぞれガウス分布の KL ダイバージェンスの期待値計算によって計算できる．第三項と第四項については，Concrete distribution の対数尤度比を期待値計算する．

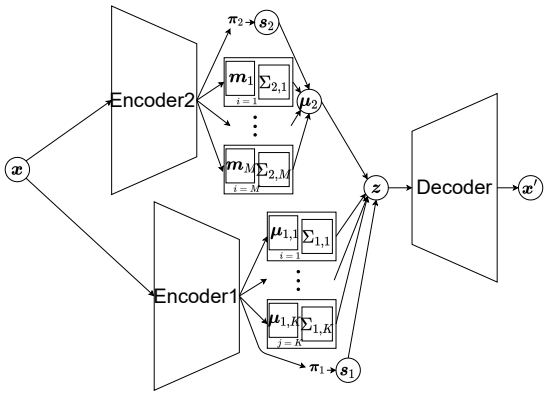

図 1 The structure of proposed VAE which consits of two encoders and a decoder.

2.3 教師データ・事前分布の与え方

本研究で用いる VAE のエンコーダ部分を図 2 に示す．図 1 における Encoder1 と Encoder2 がそれぞれ図 2 の構造を持つ．各エンコーダはそれぞれの各クラスの平均ベクトルと共分散行列そしてカテゴリカル分布のパラメータ $\boldsymbol{\pi}$ を出力する．この $\boldsymbol{\pi}$ は ArcFace[6] と呼ばれる分類器の出力から得る．この $\boldsymbol{\pi}$ の推論つまり事後分布 $q_\phi(\boldsymbol{s}_1|\boldsymbol{x}), q_\phi(\boldsymbol{s}_2|\boldsymbol{x})$ の推論がクラス分類を表す．

VAE は教師あり学習の枠組みではなく，クラス分類問題を解くためには教師データを与える必要がある．本研究では，Concrete distribution の事前分布として教師データを与える．具体的には，以下のように事前分布の各パラメータを与える．

- カテゴリ潜在変数の事前分布 $p(\boldsymbol{s}_1), p(\boldsymbol{s}_2)$ のパラメータ $\boldsymbol{\pi}_{\text{prior},1}, \boldsymbol{\pi}_{\text{prior},2}$ をそれぞれ式 (5),(6) で与える．

$$\boldsymbol{\pi}_{\text{prior},1} = \text{Softmax}(\boldsymbol{t}_1/T) \tag{5}$$

$$\boldsymbol{\pi}_{\text{prior},2} = \text{Softmax}(\boldsymbol{t}_2/T) \tag{6}$$

  ここで，$\boldsymbol{t}_1$ は人物ラベルの教師データ，$\boldsymbol{t}_2$ は方向ラベルの教師データを表すベクトルである．また，T は Softmax 関数の温度パラメータである．
- 条件つき分布 $p(\boldsymbol{\mu}_2|\boldsymbol{s}_2), p(\boldsymbol{z}|\boldsymbol{\mu}_1,\boldsymbol{s}_2)$ のパラメータは以下のように与える．ArcFace の識別層の重みベクトルが各クラスを代表するベクトルになる [6] ことを利用して，各クラスの平均として人物・方向それぞれに対応する ArcFace（分類器）の重みベクトルを与える．また共分散行列として通常の VAE と同様に単位行列を与える．

3. 骨格データの取得

本章では，識別に用いる骨格データの取得方法を述べる．天井に設置された魚眼カメラから動画を撮影し，一般物体認識手法である YOLO[7] を用いて人物領域を取得する．YOLO は魚眼画像を用いて学習を行っており，魚眼画像特有の歪んだ画像から人物領域を取得できる．さらに得られた人物領域について，正像変換を用いて魚眼画像を透視投影画像に変換する．最後に，変換された歪みの無い人物領域画像から深層学習を用いた姿勢推定手法である OpenPose[8] を用いて骨格座標を取得する．骨格座標は首の座標が原点になるよう平行移動し，首から腰の左右の中心までの長さで全体を正規化する．骨格座標データは，時系列方向に並べることで図 2 に示すように x 座標 y 座標 2 チャンネルの 2 次元データにしてからエンコーダの CNN に入力される．

4. 評価実験

本実験では，図 3 に示す二つのルート，4 つの方向について成人男性 5 人が歩行した動画から取得した骨格データを用いて，人物ラベル・方向ラベルのマルチラベルクラス分類を行った．一人当たり合計で 3000 フレームの骨格データを訓練データ，1800 フレームの骨格データをテストデータに用いた．識別に利用するフレーム長は 10，20 の 2 パターンで学習を行い評価した．学習は以下のモデルすべてそれぞれ 5000 エポック行

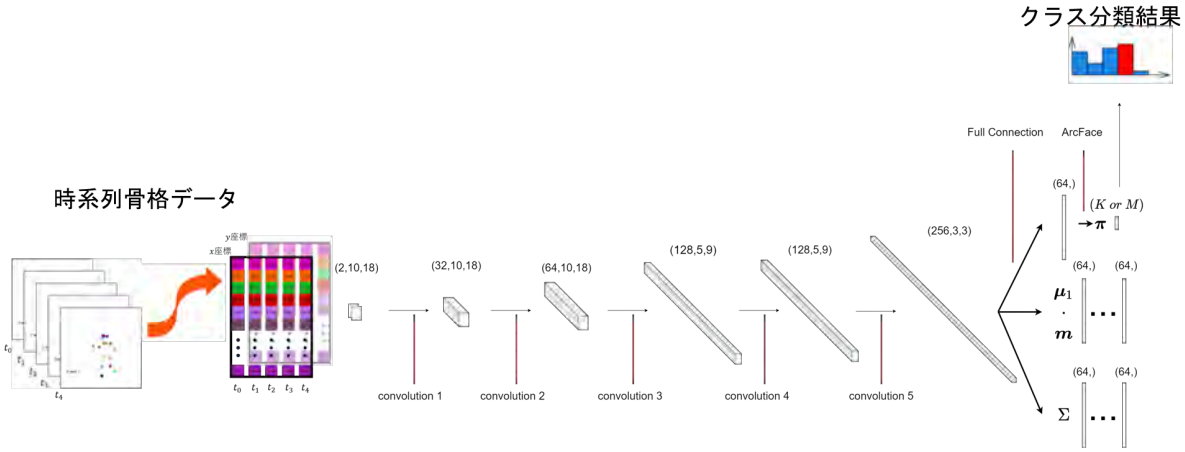

図 2　The structure of input data and encoder.

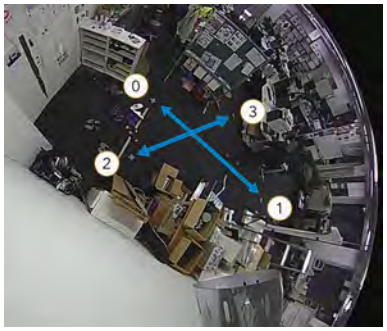

図 3 The route which subjects go through in the experiment.

い，学習係数は初期値 0.0001 として減衰させ，Gumbel Softmax-trick と事前分布の温度パラメータ T は初期値 1.0 として減衰させた．これら学習係数と温度パラメータの初期値の決定とそれらの減衰は実験的に行った．

クラス分類実験について，提案手法ではデータを入力したときの近似事後分布のパラメータ $\boldsymbol{\pi}_1$（人物ラベルの推論結果）によってクラス分類を行った．また，提案手法と比較するモデルとして，以下のモデルを用意した．

- 人物ラベルと方向ラベルのクロスエントロピー誤差で学習した識別モデル
- 人物ラベルのクロスエントロピー誤差で学習した識別モデル

以上のモデルと提案手法を用いて，人物クラス分類性能を accuracy で評価した．

歩容認証などの生体認証では，データベース上に登録された特徴と取得された特徴を比較することで本人認証と個人認証を行うことが多く，本システムでも認証手法としての利用を想定し，提案手法を特徴抽出器として利用したときの特徴抽出性能についても評価を行った．具体的には，提案手法について以下の特徴を利用して k 近傍法での識別を行った．

1. Encoder1 における ArcFace(人物クラス分類器) への入力ベクトル
2. Encoder1 と Encoder2 における ArcFace(人物クラス分類器と方向クラス分類器) への入力ベクトルを結合したもの
3. 近似事後分布からサンプリングされた潜在変数 $\boldsymbol{z}$
4. 以下で与えられる潜在変数 $\boldsymbol{z}$ に対応する平均ベクトル $\boldsymbol{\mu}_{gmm1}$

$$\boldsymbol{\mu}_{gmm1} = \left\{\sum_{i=1}^{K} s_{1,i}\Sigma_{1,i}^{-1}\right\}^{-1}\left\{\sum_{i=1}^{K} s_{1,i}\Sigma_{1,i}^{-1}(\boldsymbol{\mu}_{1,i}+\boldsymbol{\mu}_2)\right\}$$

5. $\boldsymbol{\mu}_{gmm1}$ について，$\boldsymbol{\mu}_2 = 0$ として人物ラベル成分のみ抽出したもの

ここで，$\boldsymbol{\mu}_{gmm1}$ は，近似事後分布における混合ガウス分布のうち，カテゴリ潜在変数が得られたときの条件付き分布を一つのガウス分布に変形したときの平均ベクトルを意味する．本実験では k=5 とした k 近傍法の識別結果の accuracy で評価した．

4.1　実験結果

表 1 に，近似事後分布 $q_\phi(\boldsymbol{s}_1|\boldsymbol{x})$ のパラメータ $\boldsymbol{\pi}_1$ による人物クラス分類つまり，人物識別器の出力の分類結果を示す．10,20 の各フレーム長について最良の分類結果を太字で示す．提案手法である追加ラベルを組み込んだ生成モデルによるクラス分類が，人物クラスのみの識別モデルやマルチラベルの識別モデルより良い性能を示すことがわかる．

また，表 2 に各モデルから出力される特徴を用いた k 近傍法によるクラス分類結果を示す．こちらも各フレーム長について最良の分類結果を太字で示す．提案手法によって得られた特徴，特に $\boldsymbol{\mu}_{gmm1}$ による識別性能が他のモデルより良い識別性能を示している．これは人物ラベル・方向ラベル二つのラベルを考慮した潜在変数の空間が，人物識別に有効であることを意味する．一方 $\boldsymbol{\mu}_{gmm1}$ の方がその値からサンプリングされる $\boldsymbol{z}$ よりも識別性能が高いのは，近似事後分布の共分散によって各クラス間で重複する部分が生じるためだと考えられる．

5.　おわりに

本研究では，画像から取得された時系列骨格データによる人物識別手法として，人物ラベルと方向ラベルのマルチラベルの深層生成モデルによるクラス分類を

表 1　Accuracy(%) of person classification result.

model	timestep [frame] 10	20
識別モデル (マルチラベル)	92.5	94.5
識別モデル	92.2	94.8
提案手法	**93.2**	**95.4**

表 2 Accuracy(%) of person classification result with kNN.

model	feature	timestep [frame] 10	20
識別モデル	人物識別器	91.7	94.4
識別モデル (マルチラベル)	人物識別器	92.0	94.6
	人物・方向識別器	93.2	95.5
提案手法	人物識別器	92.2	94.1
	人物・方向識別器	92.7	94.4
	潜在変数 $\boldsymbol{z}$	90.2	88.2
	$\boldsymbol{\mu}_{gmm1}$	**93.4**	95.7
	$\boldsymbol{\mu}_{gmm1}(\boldsymbol{\mu}_2 = \mathbf{0})$	93.2	**95.8**

提案し，クラス分類・特徴抽出性能ともに有効であることを確認した．本研究では人物ラベルと方向ラベルのみ利用しており，手荷物の有無や服装などといった歩容特徴に大きく影響する因子が考慮できていないため，今後はより多数のラベルを用いた生成モデルによる分類を目指す．また，本研究で用いた生成モデルは階層的で複雑であり，学習に時間がかかったり学習が不安定である．今後はシンプルなモデルや分散の小さい近似計算を用いることで，学習の安定化を検討する．

参　考　文　献

[1] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, “Multiple views gait recognition using view transformation model based on optimized gait energy image,” *International Conference on Computer Vision Workshops*, pp. 1058-1064, 2009.

[2] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, “Geinet: View-invariant gait recognition using a convolutional neural network,” *International Conference on Biometrics*, pp 1-8, 2016.

[3] 守脇幸佑，村松大吾，武村紀子，八木康史，“追加ラベルを組み込んだ歩容特徴抽出器”，電子情報通信学会技術報告，Vol. 119, No. 214, pp. 31-35, 2019.

[4] C.J. Maddison, A. Mnih, and Y.W. Teh, “The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables,” *Internatinal Conference on Learning Representations*, 2017.

[5] D.P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *Internatinal Conference on Learning Representations*, 2014.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” *Conference on Computer Vision and Pattern Recognition*, pp. 4690-4699, 2019.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[8] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, “Real-time multi-person 2d pose estimation using part affinity fields,” *Conference on Computer Vison and Pattern Recognition*, pp. 7291-7299, 2017.