差分ステレオを用いた複数人物のセグメンテーション

Segmentation of Multiple Human Using Subtraction Stereo

○学　生形　徹（中央大）　モロ　アレッサンドロ（University of Trieste /JST CREST）
星川　佑磨（中央大/JST CREST）　学　有江　誠（中央大）
正　寺林　賢司（中央大/JST CREST）　正　梅田　和昇（中央大/JST CREST）

Toru UBUKATA, Chuo University, ubukata@sensor.mech.chuo-u.ac.jp
Alessandro MORO, University of Trieste / CREST, JST, alessandro.moro@stud.units.it
Yuma HOSHIKAWA, Chuo University / CREST, JST, hoshika@sensor.mech.chuo-u.ac.jp
Makoto ARIE, Chuo University, arie@sensor.mech.chuo-u.ac.jp
Kenji TERABAYASHI, Chuo University / CREST, JST, terabayashi@mech.chuo-u.ac.jp
Kazunori UMEDA, Chuo University / CREST, JST, umeda@mech.chuo-u.ac.jp

In this paper, we propose a method for segmentation of multiple human in a projection plane. Our algorithm requires a stereo camera system called Subtraction Stereo, which extracts foreground information with a fixed stereo camera. The main contribution of this paper is how the image sequences that include partial occlusion of the foreground objects can be accurately segmented using mean shift clustering in real-time processing. The proposed method is suitable for inside a medium-sized environment, such as a room and an entrance. Finally, we try to segment the image sequences that include partial and full occlusion and show the accuracy of the proposed method.

***Key Words*:** Stereo Camera, Surveillance Camera, Segmentation, Mean Shift Clustering

1. 緒言

近年，防犯意識の高まりと共に街中に多くの監視カメラが設置されるようになってきている．監視カメラの目的の多くは不審者の監視や人流計測による安全管理などであるが，これらを実現するためには多くの人手と人件費が必要となってくる．そのため，監視カメラ用途を想定し，カメラ画像から自動で人物検出やトラッキングを行うシステムが多く研究されている [1, 2, 3, 4].

監視カメラの映像ではカメラに対して人物が重なることによってオクルージョン (遮蔽) を生じる場面がしばしば見られる．このオクルージョン問題は人物を検出する際に大きな障害となるため，多くの研究者達がこの問題に取り組んでおり，そのための手法はいくつかのタイプに分けることができる．1 つ目は機械学習を応用した，特徴量ベースの物体識別である．Dalal らは Histograms of Oriented Gradients (HOG) 特徴量を識別器にかけることによってオクルージョンに対応できる人物検出手法を構築した [5]. 2 つ目は最適化手法を用いたオブジェクトセグメンテーションである．Chaohui らは Markov Random Field (MRF) という確率モデルを用いることで，厳しいオクルージョンに対しても正確にセグメンテーションを行うことを可能にした [3]. これら 2 つの手法はオクルージョン問題に対して有効な手法であるが，処理時間が多く実時間処理に不向きであることから監視カメラ用途には適さない．そこで，本研究で注目した手法が，3 次元情報を元にセグメンテーションを行うものである．Bahadori らはステレオカメラから得られる 3 次元情報を投影することで実時間処理でのセグメンテーションを可能にしたが，人物間の距離が小さい場合において精度が著しく低下してしまうという問題があった [1].

この問題に対処するため, 本研究では Bahadori らの手法 [1] に (1) 投影点のヒストグラム化，(2) ヒストグラム上での Mean shift クラスタリングという 2 つの改善点を加えることで実時間処理でより正確なセグメンテーション手法を構築する．Mean shift クラスタリングとは，トラッキング手法としてよく用いられる Mean shift 法 [6] をクラスタリング手法として応用したものである．これによって，従来では利用していなかった投影点の密度を利用してセグメンテーションを行うことができる．

本稿の構成は以下の通りである．まず，2 章において本研究でのオブジェクト検出の基盤となっている差分ステレオ [7] と影検出について述べる．3 章では，3 次元情報を利用した複数人物のセグメンテーション手法について詳しく述べる．続く 4 章では，実環境での実験を通して，提案手法の有用性を Bahadori らの手法 [1] と比較評価することによって示す．最後に 5 章で結論と今後の展望について述べる．

2. 前景検出

2.1 差分ステレオ

本研究では画像中のオブジェクトを検出する手法として，固定されたステレオカメラを用いた，差分ステレオ [7] と呼ばれるシステムを用いている．差分ステレオの基本アルゴリズムを図 1 に示す．通常のステレオカメラでは，左右カメラの画像をマッチングすることで視差画像を得る．これに対し，差分ステレオでは左右カメラそれぞれで，まず背景差分によって前景領域を抽出し，その後に抽出された前景領域をマッチングすることで距離情報を得る．このように，マッチングする領域を背景差分で得られた前景領域に限定することで，マッチングの誤対応と処理時間を削減することができる．差分ステレオ によって検出された前景領域を図 2 (b) に青色で示す.

2.2 影検出

差分ステレオによって検出された前景領域には，オブジェクトの影となる領域も含まれてしまう．そのため，オブジェクトを正確に検出するためには，影となる領域を前景領域から除去する必要がある．

画像座標(x, y)における輝度値を$I(x, y)$とし，背景画像における同位置の輝度値を$I'(x, y)$とすると，影を判定する評価関数は次式で表わされる．

$$\theta_{(t+1,x,y)} = \begin{cases} \alpha\Psi_{(x,y)} + \beta\Lambda_{(x,y)} + (1-\alpha-\beta)\theta_{(t,x,y)}, \\ \qquad if \ \dfrac{I_{(x,y)}}{\eta} < I'_{(x,y)} \\ \infty, otherwise \end{cases} \quad (1)$$

ここで，θは影と判断するためのスコアを表し，θが閾値以下となる画素を影と判断して前景領域から除去する．Ψは前景の近傍画値と背景の近傍画素値の相違度を表し，Λ は前景の色相と背景の色相の相違度を表す．また，α，β，η はそれぞれの項に対して重みを与える定数である．

今回用いた影検出手法の詳細は文献 [8] にて述べられており，その手法を用いて影検出を行った結果を図 2 (c) に示す．図中において，影と認識された領域を緑色で表わし，前景領域から除去している．

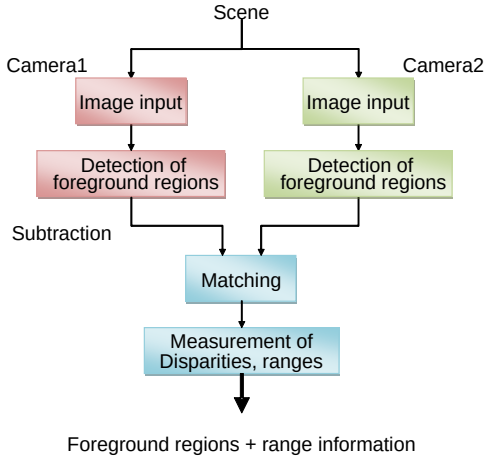
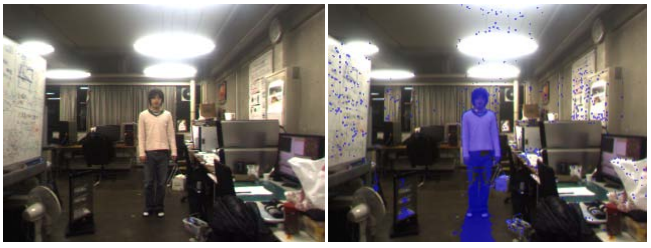
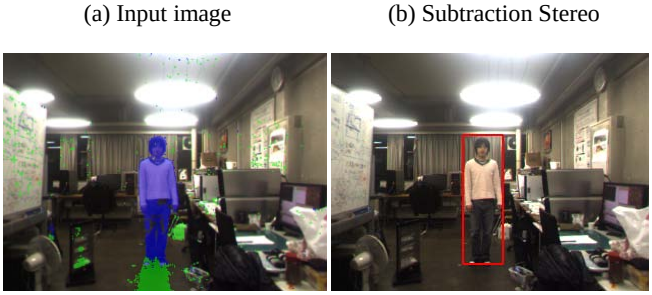


Fig.1 Flow of the subtraction stereo



(a) Input image　(b) Subtraction Stereo



(c) Shadow detection　(d) Object detection

Fig.2 Flow of the object detection

3. 投影平面でのセグメンテーション

最終的に検出された前景画素は 8 近傍ラベリングにより，連結画素ごとにひとつのオブジェクトとして認識される．オブジェクトの認識結果を図 2 (d) に赤い矩形で表す．ここで問題となるのが，画面中に複数の人物が存在し，オクルージョンが発生する場合，複数の人物を 1 つのオブジェクトとして認識してしまうことである(図 3(a))．このような場面において，本研究では前景の 3 次元情報を元にセグメンテーションを行う．

カメラの位置，姿勢が既知である時，ステレオカメラから得られるカメラ座標系での 3 次元情報を世界座標系へ変換することができる．本手法では，前景領域の距離情報を世界座標系の平面に投影することで複数人物のセグメンテーションを行い，より正確なセグメンテーションを可能にするために Mean shift クラスタリングを用いる．

3.1 投影平面

前景から検出されたオブジェクトの総数を n とし，それぞれのオブジェクトを$\{B_i | i = 0, \ldots, n\}$と定義する．まず，検出されたオブジェクトの 3 次元情報を世界座標系 X-Y 平面(地面)に投影する．ここで言う世界座標系 X-Y 平面とは上から人物を俯瞰した面を表わし，今後この平面を投影平面と呼ぶ．図 3 (c) に図 3 (a) の前景領域を投影平面に投影した様子を示す．ここで，画像座標(x, y)における画素の投影平面上での位置を$p(x, y)$と定義する．

投影平面での処理を簡易にするため，投影点の広がりを考慮に入れ，投影平面に 5cm × 5cm のセルを構成する．そして，このセル中に何点の投影点が投影されているかをカウントすることにより，次式で表わされるヒストグラムを構成する．

$$H(c) = \left\{ \sum N_{(x,y,c)} \,\middle|\, \forall (x, y) \in B_i \right\} \quad (2)$$

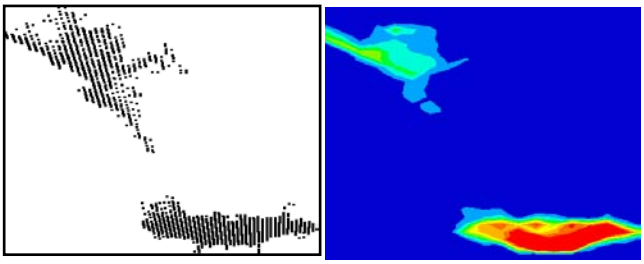
$$where \quad N_{(x,y,c)} = \begin{cases} 1, if\ p(x, y) \subseteq c \\ 0, otherwise \end{cases}$$

ここで，c は投影平面上のセルひとつの領域を表す．

ヒストグラムを構成した後，ヒストグラムの頻度が閾値以上のセルを対象に，8 近傍ラベリングによって連結領域を検出する．今後，投影平面上でのセル連結領域を *Projected blob* と呼ぶ．図 3 (d) に図 3 (c) から得られた *Projected blob* を示し，ヒストグラムの頻度を赤(高)から青(低)のグラデーションで表す．ここで，*Projected blob* の総数を m とし，各 *Projected blob* を$\{PB_j | (j = 0, \ldots, m)\}$と定義する．セル位置と画像座標には相関関係があるので，*Projected blob* ごとに個人と認識することで画像平面上においても図 3 (b) に示すようなセグメンテーションが可能である．

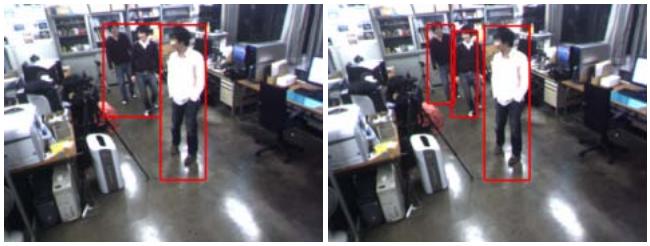
しかし，前述したように人物間の距離が小さくなる場合，*Projected blob* を構成するだけではセグメンテーションが困難である．人物間距離が小さい場合の例を図 4 (a) の左側 2 人組に示し，左側 2 人組の *Projected blob* を図 4 (c) に示す．図 4 (c) から見てとれるように，人物間距離が小さい時に精度が低下するのは，個人ごとに検出されるはずの *Projected blob* が連結してしまうことが原因となる．そこで，本研究ではこのような場合において，次章で述べられる Mean shift クラスタリングを用いる．

(a) Error of object detection (b) Result of the projection plane

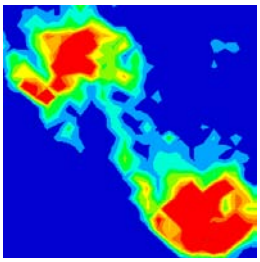


(c) Projected points (d) Projected blobs

Fig.3 Comparison with and without a projection plane



(a) Bahadori's method [1] (b) After the mean shift clustering



(c) *Projected blobs* in the left bounding box in (a)

Fig.4 Comparison with and without mean shift clustering

3.2 Mean shift クラスタリング

Mean shift クラスタリングとは教師なしクラスタリング手法の一種である．投影点をデータ群と見立てることで，反復計算から投影点の密度推定を行うことができることから，その密度情報を利用してクラスタリングを行うことが可能である．本手法では投影点をデータ群として直接 Mean shift クラスタリングを行うのではなく，3.1 節で構成したヒストグラム上で行う．これによって，投影点のもつ距離情報を疎に扱うことができるため，処理時間を削減することができる．

Projected blob での任意のセルにおける位置ベクトルをP_cとすると，Mean shift ベクトル$m(v)$は次式で表わされる．

$$m(v) = \frac{\sum_{c \in rectangle} P_c H(c)}{\sum_{c \in rectangle} H(c)} - v \qquad (3)$$

ここで，v は重心位置である．$H(c)$は式(2)に示すように投影点の密度を表わしているので，Mean shift ベクトルを算出する際の重みとして用いられる．実時間処理を可能にするために，本手法では *rectangle* (矩形) をカーネル (重心計算を行う範囲) として用いる．*rectangle* の大きさは任意の距離における投影点の広がりから算出しており，中心は v に位置している．また，カーネルの数は *Projected blob* の大きさに応じて変化させており，初期位置は *Projected blob* 内に一様に配置する．Mean shift の結果は以下の反復計算によって得られる．

1. カーネル範囲内で式(3)を算出
2. カーネルの中心を$v^{t+1} = v^t + m(v^t)$へ移動

この反復計算は密度分布の勾配が 0 になる場所に収束することが保証されており，Mean shift ベクトルの移動量が十分に小さくなった時に止められる．

全てのPB_j上で Mean shift の処理が行われた後，同位置または近隣のセルに収束したカーネルが統合される．この統合されたカーネルの範囲内にあるセルごとにクラスタリングを行う．以上のクラスタリング手法を，図 4 (a) と同様の画像に用いた結果を図 4 (b) に示す．

4. 実験結果

本章では前章までに説明した提案手法の有用性を評価するために，オクルージョンの頻出する 2 シーンで実験を行った．実験時のステレオカメラは Bumblebee2 (Point Grey Research, VGA) を使用し，30 [fps] で保存した動画に対して処理を行った．処理に用いた PC は CPU が Intel Core2 Extreme (2.93 GHz), メモリが 3 GB RAM を搭載したものを使用した．

提案手法を評価するにあたり，Bahadori らの手法 [1] を構築し，提案手法との比較を行う．提案手法を用いた実験結果を図 5 に示し，それぞれの動画で 1,000 フレーム分を本手法と Bahadori らの手法 [1] とで評価し，比較した結果を表 1 に示す．ここで，T.Pos.は正しい検出，F.Neg.は未検出，F.Pos.は誤検出を表す．図 5 から見てとれるように，本手法では人物同士が近くにいる場面においてもセグメンテーションが可能であり，その結果，表 1 のように Bahadori らの手法[1]より高い精度が得られたと考えられる．また，実験時の平均処理時間は Room が 56[ms]，Elevator が 66[ms]となっている．オフラインでの処理は動画を転送して読み込む処理が余分にかかることから，実時間処理が可能であると考えられる．

提案手法での失敗例を図 6 に示す．図 6 (a) では後方にいる人物が検出されていないのがわかる．これは，オクルージョンが厳しすぎて，投影する距離情報が不足することによって生じる．正確に検出するためには，人物の半身程度が見えていることが条件となる．また，図 6 (b) では人物が見えているにも関わらず，右の人物が検出されていない．これは，Mean shift クラスタリングの際にヒストグラムの局所最大値にカーネルが収束してしまったことが原因である．

提案手法は距離情報をもとにセグメンテーションを行っていることから，距離計測精度に非常に依存している手法ということが言える．一方，ステレオカメラの距離計測精度は距離の 2 乗に比例して誤差が大きくなるため，提案手法では遠距離でのセグメンテーションが困難である．計測対象が遠い場合の検出結果を図 7 に示す．そのため，提案手法の実用的な計測範囲は 10m 程度の室内や廊下などの環境に限定される．

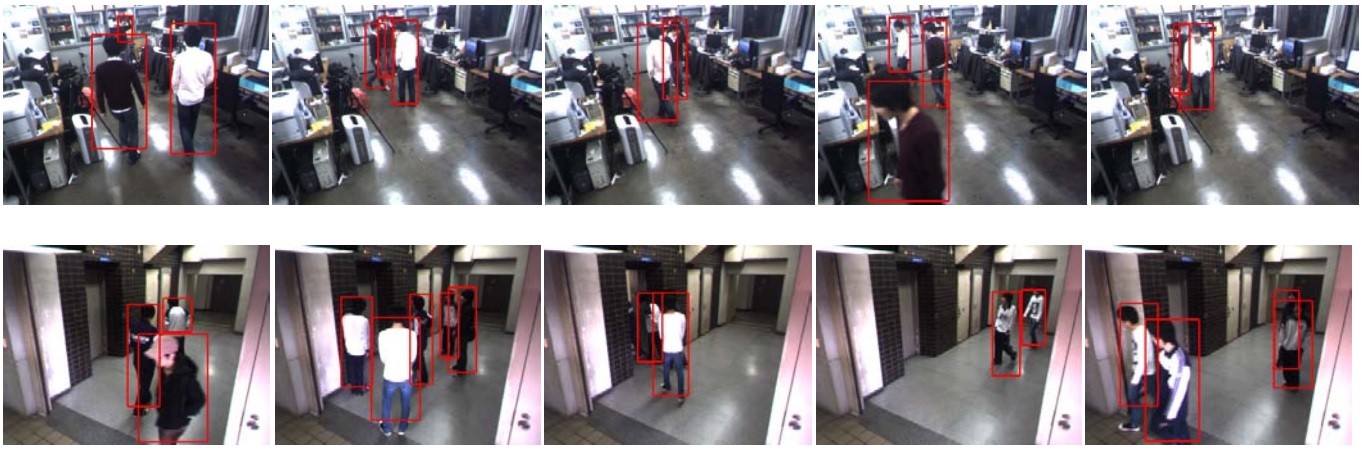
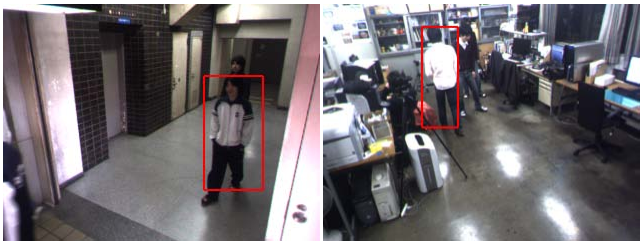


Fig.5 Experimental results: in the room (above); in front of the elevator (below)



(a) (b)

Fig.6 Examples of error scenes

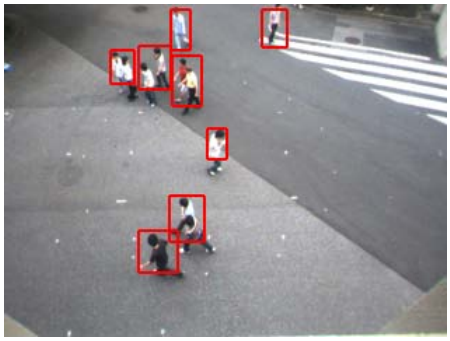


Fig.7 Example of the detection in a far place

5. 結論

本稿では，差分ステレオより得られる距離情報を投影平面上で処理することで，複数人物のセグメンテーションを行う手法を提案した．投影平面へ投影された距離情報は Mean shift クラスタリングを用いることで，より正確なセグメンテーションが可能となり，実験を通してその有用性が確認された．また，ステレオカメラの距離計測精度より，提案手法の計測範囲の限界を示した．

今後は，計測対象が遠い環境においても人物の検出を行えるように，特徴量を用いた機械学習での人物検出手法と統合していく．また，既存のトラッキング手法 [9] と組み合わせることで，より汎用性の高い監視カメラシステムを構築していく予定である．

Table 1 Evaluation result

Sequence	T. Pos. [%]	F. Neg. [%]	F. Pos. [%]
Room	88.9	11.1	0.2
Room [1]	80.7	19.3	4.8
Elevator	88.4	11.6	1.5
Elevator [1]	81.3	18.7	2.4

文 献

[1] Bahadori S., et al., “Real-time people localization and tracking through fixed stereo vision”, Applied Intelligence, Vol.26, No.2, pp.83-97, 2007.

[2] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, “Robust Tracking-by-Detection using a Detector Confidence Particle Filter” In *Proc. ICCV*, pp. 1515–1512,, 2009.

[3] W. Chaohui., et al., “Segmentation, Ordering and Multi-Object Tracking using Graphical Models”, In *Proc. ICCV*, pp. 747–754, 2009.

[4] Y. Ma, S. Worrall and A. M. Kondoz, “Depth Assisted Visual Tracking”, In *Proc. WIAMIS*, pp. 157–160, 2009.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", In *Proc. CVPR*, CA, USA, pp. 886 – 893, 2005.

[6] D. Comaniciu, V. Ramesh, P. Meer, “Real-time tracking of non-rigid objects using mean shift”, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, pp.142–149. 2000.

[7] Umeda K., et al., “Subtraction Stereo -A Stereo Camera System That Focuses On Moving Regions -”, *Proc. of SPIE-IS&T Electronic Imaging*, Vol.7239 Three-Dimensional Imaging Metrology, 723908, 2009.

[8] A. Moro, et al., “Auto-adaptive threshold and shadow detection approaches for pedestrians detection”, In *Proc AWSVCI*, pp. 9-12, 2009.

[9] Hoshikawa Y, et al., “Human Tracking Using Subtraction Stereo and Color Information”, In *Proc AWSVCI*, pp. 5-8, 2009.