修士論文

映像検索に向けた
人物の着衣推定

２０２１年度

武 藤　良

明星大学大学院
情報学研究科情報学専攻

20MJ-006

目次

表目次

図目次

概要

本研究では，大規模映像データベースから特定の着衣の情報を含む映像を検索する映像検索システムを提案する．

大規模映像データから特定の映像を検索するためには，映像中の物体，背景などの幅広い情報を認識する必要がある．画像から幅広い特徴を得る手法として，言語と視覚的特徴をマルチモーダルに学習する言語・画像埋め込み手法が存在する．しかし，言語・画像埋め込み手法では画像中の物体の詳細な属性の認識が難しい．

提案手法では言語・画像埋め込み手法の欠点を補う大規模映像データベースから詳細な人物の着衣を検索するシステムを提案する．提案するシステムは，言語・画像埋め込み技術を用いて映像から特徴量を抽出する言語-画像類似度算出モジュールと，映像中の人物の詳細な着衣の特徴を抽出する着衣・色識別モジュールの 2 つのモジュールで構築される．

言語-画像類似度算出モジュールでは，言語・画像埋め込み手法のエンコーダに検索クエリ・映像データを入力し，共通空間のベクトルを出力する．出力された特徴ベクトル同士の類似度を算出し，検索クエリに沿った映像データが検索結果上位となるような検索スコアを算出する．

着衣・色識別モジュールでは，映像中の人物・着衣の領域，人物の関節点検出の複数の手法を用いて着衣のカテゴリ情報の推定とその色情報の抽出を行う．本モジュールではインスタンスセグメンテーションを用いて人物領域と着衣領域・着衣カテゴリ，姿勢推定を用いて人物の上半身／下半身領域を検出し，各領域から着衣の色情報を抽出する．着衣のカテゴリ情報と色情報から，検索クエリに沿った着衣が写った画像が検索結果上位になるような検索スコアを出力する．

提案手法を用いて，大規模映像データベースから着衣が関係する検索クエリを用いて検索実験を行った．その結果，着衣・色識別モジュールの検索スコアを補助的に言語-画像類似度算出モジュールのスコアに統合することで検索精度の向上が見られた．

また，各モジュール単体の検索スコアのみを用いて同様の検索実験を行ったところ，検索結果上位の画像に検索対象に設定した着衣が確認できた．しかし，着衣・色識別モジュールのみでは検索クエリに含まれる着衣以外の要素についての解析できず，また言語-画像類似度算出モジュールのみでは詳細な着衣の検索が行えていないかった．このことから，両モジュールを組み合わせることで詳細な映像検索に有効であることがわかった．

1　序論

近年，動画投稿サイトや SNS の発展により，インターネット上にアップロードされる映像データの数は飛躍的に増加すると考えられる．また，高品質，低遅延であることを利用した防犯カメラの映像データを用いたリアルタイムなセキュリティシステムも登場している．これらの環境から，大規模な映像データベースからユーザーが求める特定の映像，シーンを検索することができる映像検索システムが注目されている．特に，映像中の人物の属性を識別した映像を検索することができれば，防犯カメラの映像から人物の特定や，映像中の人物の着衣からファッショントレンドの推定，動画投稿サイトから見たい動画の検索など，幅広い分野に活用できることが考えられる．

ユーザが指定した属性をとらえた人物を含む映像の検索を実現する手法として，膨大な画像-言語間の関係性を学習した画像・言語埋め込みモデルが提案されている．画像・言語埋め込みモデルは，画像と言語の特徴をベクトルで表現することができる手法である．映像データのキーフレーム画像から得られる特徴ベクトルと検索クエリとなるテキストの類似度を計算することで，ユーザがテキストで指定した特徴に一致した映像データを検索することができる．

また，映像中の人物の属性を推定する別の手法として，インスタンスセグメンテーションや姿勢推定などの手法も提案されている．これらの手法は人物の領域を検出することができるため，映像中の人物一人あたりの詳細な属性推定に活用することができる．

本研究では，言語-画像間の関係性を学習した画像・言語埋め込み手法と，インスタンスセグメンテーション・姿勢推定を組み合わることで，画像・言語間の幅広い特徴を捉える能力と，詳細な人物の属性推定を両立させた映像検索システムを提案する．

2　関連研究

2.1　文章からの映像検索：画像・言語埋め込み手法

大規模映像データベースからユーザーが指定する映像データ検索する映像検索の研究は，長年行われている．米国国立標準技術研究所（NIST）が主催する映像検索ベンチマークであるTRECVID [1] では，2016 年から新たに Ad-hoc Video Search（AVS）タスクが開始された．このタスクは，大規模映像データベースから提示された検索クエリ（文章）をもとに，検索クエリに合致した映像検索を行うタスクである．提示される検索クエリは，「Find shots of a man in blue jeans outdoors」といった人物の属性が関係するもの，「Find shots of train tracks during the daytime」といった背景や時間帯が関係するものなど，幅広い意味をもつ．そのため，どのような意味を持つ検索クエリが入力されても対応できるゼロショット機能をもつ映像検索システムが求められる．

入力検索クエリからゼロショットの映像検索を実現するシステムとして，自然言語と視覚的情報の関係性を学習する画像・言語埋め込みの手法が提案されている．Object-Semantics Aligned Pre-training（Oscar）[2] は，画像から物体検出で得られた物体の情報をタグとして画像-言語間の関係性の学習の補助に用いる．視覚的な位置情報をもつタグを学習に用いることで，画像と文章の特徴の位置合わせを考慮した画像-文章の関係性を学習している．Contrastive Language-Image Pre-training（以下，CLIP とする）[3] は，画像とその画像を説明する文章のペアを入力として，画像から特徴ベクトルを出力するイメージエンコーダと，文章から特徴ベクトルを出力するテキストエンコーダからなる．正しい組み合わせの画像-文章のペアの特徴ベクトルの類似度が高くなるような学習を行うことで，入力された画像・文章から共通の特徴を表すベクトルを算出することができるようなイメージエンコーダ・テキストエンコーダを訓練する．CLIP はインターネット上の膨大な数の画像-文章のペアを学習しているため，入力画像・文章から一般的な幅広い意味をもつ特徴ベクトルを算出することができる．しかし，CLIP は画像中の物体の詳細な属性（車の車種，花の種類など）を識別する能力が低いことが課題にあげられている．

2.2　人物の検出・着衣識別の手法

映像中の詳細な人物の属性の解析を行うためには，物体検出，姿勢推定，領域抽出等の技術が活用できる．例えば，映像中の物体の分類，領域の検出を行う手法として，Mask R-CNN [4]，YOLACT [5]，SOLO [6]，SOLOv2 [7] などのインスタンスセグメンテーションの手法が存在する．インスタンスセグメンテーションは，隣接した同種類の物体を区別しながらピクセルにクラスラベルを関連付けることができる．この手法を用いることで，映像中の人物，着衣の識別，領域の検出を行うことができる．また，OpenPose [8] や DensePose [9] のように代表される，人

物の関節情報を検出できる姿勢推定の手法がある．姿勢推定は人物の関節点の座標を求めることができ，人物とその部位が検出できれば着衣情報の分析に活用できる．

人物の特定の領域に着目した人物の着衣推定の研究も多く行われている．PANDA [10] では，人物の領域を Poselet と呼ばれる小さな領域に切り分け，各 Poselet から畳み込みニューラルネットワークを用いて特徴抽出を行う．すべての Poselet から得られた特徴量で線形分類器を作成し，人物の着衣を含む属性を推定する．FashionNet [11] では，畳み込みニューラルネットワークによるカテゴリ・属性推定ネットワークに，着衣のランドマーク座標を推定するランドマーク推定ネットワークを組み込んだ着衣推定ネットワークである．ランドマーク周辺の特徴量を補助的に着衣の属性推定に組み込んでいる．また，実験に使用するデータセットとして，着衣カテゴリ・ランドマーク情報のアノテーションを持つ大規模ファッション画像データベース DeepFashion を提案した．また，Wenguan Wang らは，着衣のカテゴリ・属性推定とランドマーク検出の課題に対して，学習時に注目する領域を指定する 2 つの Attention Module1 を導入した [12]. この研究では，ネットワークに着衣のランドマーク領域に注目する Fashion Landmark-Aware Attention と，着衣のカテゴリ分類をするために大域的に注目する Clothiong Category-Driven Attention を組み込むことで，着衣ランドマークと着衣カテゴリの同時推定を可能としている．

これらの手法では，画像中の着衣を推定するときに，人物の部位に着目したアプローチが有効であることが示されている．

3　提案手法

本手法では，画像・言語埋め込み手法を用いた画像・言語の特徴ベクトルの類似度を算出する機能と，インスタンスセグメンテーションと姿勢推定を用いて画像中の人物の着衣の解析を行う機能を組み合わせた，大規模映像データベースから特定の着衣を身につけた人物を検索するシステムを提案する．

システム全体の概要図を図1に示す．提案するシステムは，以下の 2 つのモジュールから構成される．

- 検索クエリと映像の特徴量を比較したスコアを出力する言語-画像類似度算出モジュール
- 映像中の着衣の種類の推定とカラースコアの出力を行う着衣・色識別モジュール

各モジュールは映像キーフレームに対して検索クエリに応じた検索スコアを出力する．出力したスコアを統合し，キーフレーム画像 1 個に対する統合検索スコアを出力する．すべてのキーフレーム画像の統合検索スコアを降順で並べ，検索を行う．言語-画像類似度算出モジュールについては3.1節，着衣・色識別モジュールについては3.2節で解説する．

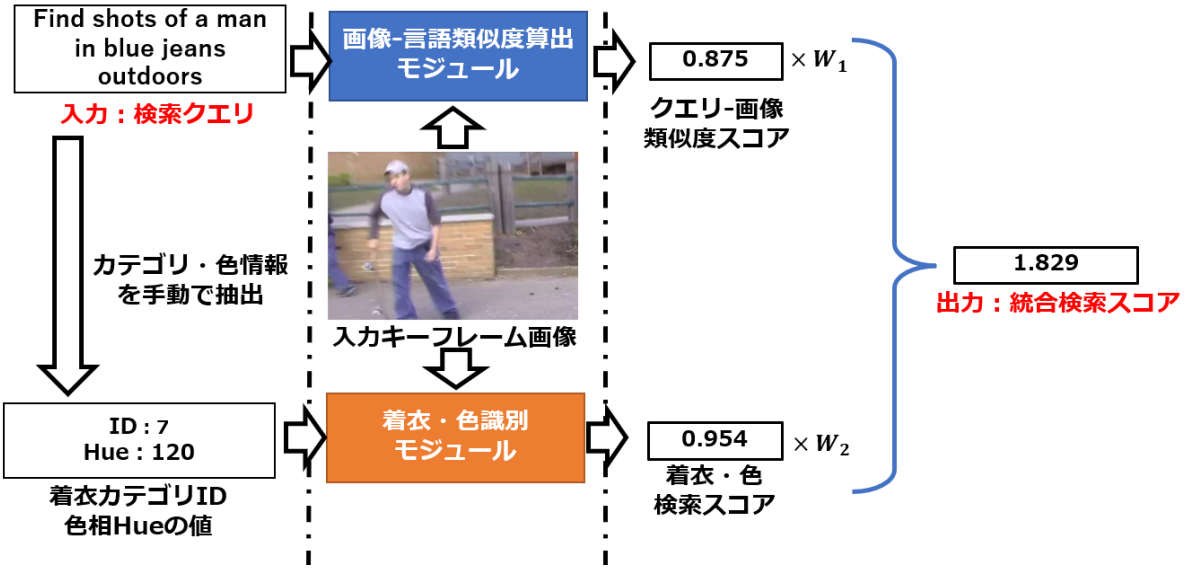

図1　システム全体の概要図．言語-画像類似度算出モジュールと着衣・色識別モジュールを用いて大規模映像データベースのすべての映像からクエリ-画像類似度スコアと着衣・色検索スコアを算出，統合する．統合された検索スコアをもとにランキングを行うことで映像検索を行う．言語-画像類似度算出モジュールの入力は検索クエリ（テキスト），着衣・色識別モジュールの入力は着衣カテゴリ ID と色相（Hue，0～180 の範囲）の値である．

3.1　言語-画像類似度算出モジュール

画像-言語類似度算出モジュールは，画像・言語埋め込み手法のテキストエンコーダとイメージエンコーダを用いて検索クエリとキーフレーム画像の特徴の類似度をクエリ-画像類似度スコアとして出力する．言語-画像類似度算出モジュールの入力は，映像データのキーフレーム画像，検索したい映像の特徴を示す検索クエリ（テキスト）である．本手法では，入力に使用する検索クエリは「Find shots of a man in blue jeans outdoors」のような，人物の着衣が関係する情報を含むものを想定している．画像-言語類似度算出モジュールの概要図を図2に示す．

言語-画像類似度算出モジュールの出力スコアは，入力された映像キーフレーム画像と検索クエリの特徴のコサイン類似度の値である．

特徴ベクトルを算出するために，入力された映像キーフレーム画像をイメージエンコーダ，検索クエリをテキストエンコーダに入力する．テキストエンコーダはテキストから，イメージエンコーダは画像から共通の空間をもつ特徴ベクトルを出力エンコーダである．両エンコーダから出力された特徴ベクトルのコサイン類似度を計算し，キーフレーム画像 1 個に対するクエリ-画像類似度スコアとして出力する．クエリ-画像類似度スコアが高いキーフレーム画像ほど，検索クエリにの内容に沿った特徴をもつ（正解データに近い）画像である．

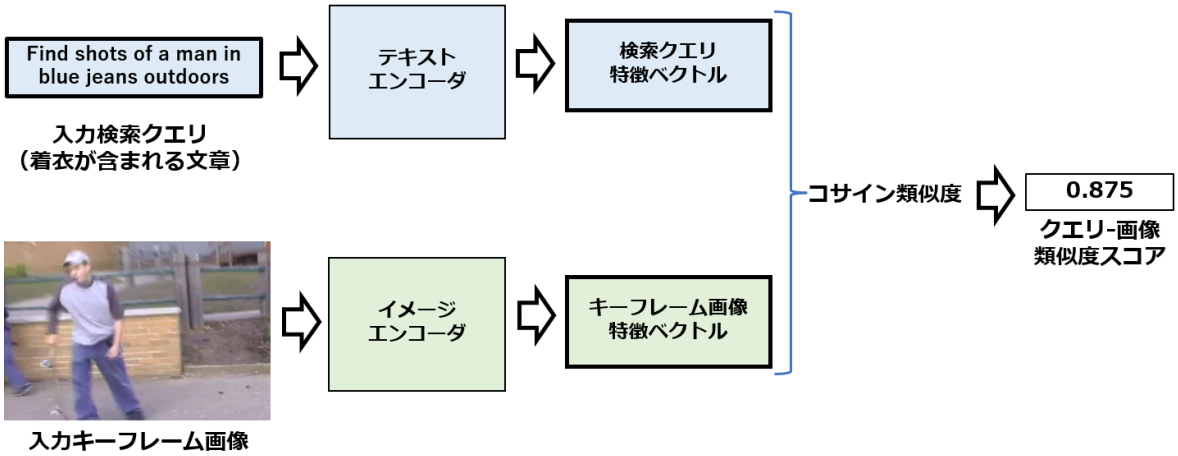

図2　言語-画像類似度算出モジュールの概要図．入力された検索クエリの特徴量を画像-言語埋め込み手法のテキストエンコーダに入力し得られた特徴ベクトルと，検索対象となる大規模映像データベースの映像のキーフレーム画像の特徴ベクトルとのコサイン類似度をクエリ-画像類似度スコアとして算出する．スコアが高い画像ほど，検索クエリの内容に近い特徴をもったキーフレーム画像である．

3.2　着衣・色識別モジュール

着衣・色識別モジュールは，インスタンスセグメンテーションと姿勢推定の技術を用いて着衣のカテゴリ・色情報を抽出，着衣スコアとして出力するモジュールである．着衣・色識別モジュールの概要図を図4に示す．

入力は，映像データのキーフレーム画像，検索対象となる着衣のカテゴリ ID と，検索対象となる着衣の色の色相の値である．カテゴリ ID と着衣の色は，言語-画像類似度算出モジュールに入力する検索クエリの内容に応じて手動で設定する．カテゴリ ID は，DeepFashion2 [13] のカテゴリ設定に従って 0〜12 の範囲で設定する．DeepFashion2 の詳細は3.2.2節で解説する．色相の値は，0〜180 の範囲で設定する．色相の値と実際の色の関係性を図3に示す．例として，「青色の長ズボン」を検索対象とする場合は，カテゴリ ID は 7（Trousers），色相の値は 120 として入力する．入力された着衣カテゴリ ID と色相の値をもとに大規模映像データベースのすべての映像から着衣・色検索スコアを出力する．

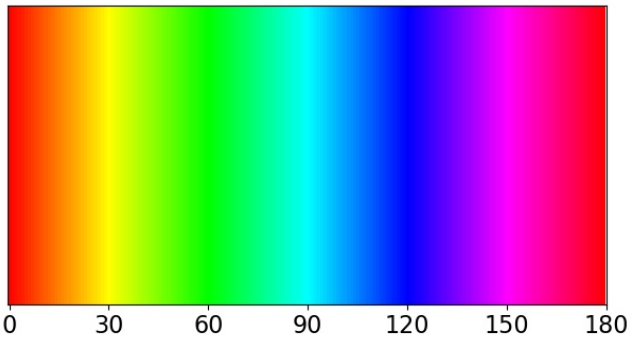

図3　色相の値 0〜180 と実際の色の関係．例として，青色の着衣を検索する場合は色相を 120 として設定する．

映像データごとの着衣・色検索スコアの算出には，映像データのキーフレーム画像からインスタンスセグメンテーションで検出した人物領域，着衣領域から抽出した色情報とその分類スコア，姿勢推定で検出した人物領域から抽出した色情報を用いる．本モジュールでは，SOLOv2 人物領域マスク，SOLOv2 着衣領域マスク，OpenPose 人物領域マスクの 3 つのマスクを生成し，着衣・色検索スコアの算出を行う．SOLOv2 人物領域マスクについて3.2.1節，SOLOv2 着衣領域マスクについて3.2.2節，OpenPose 人物領域マスクについて3.2.3節で説明する．

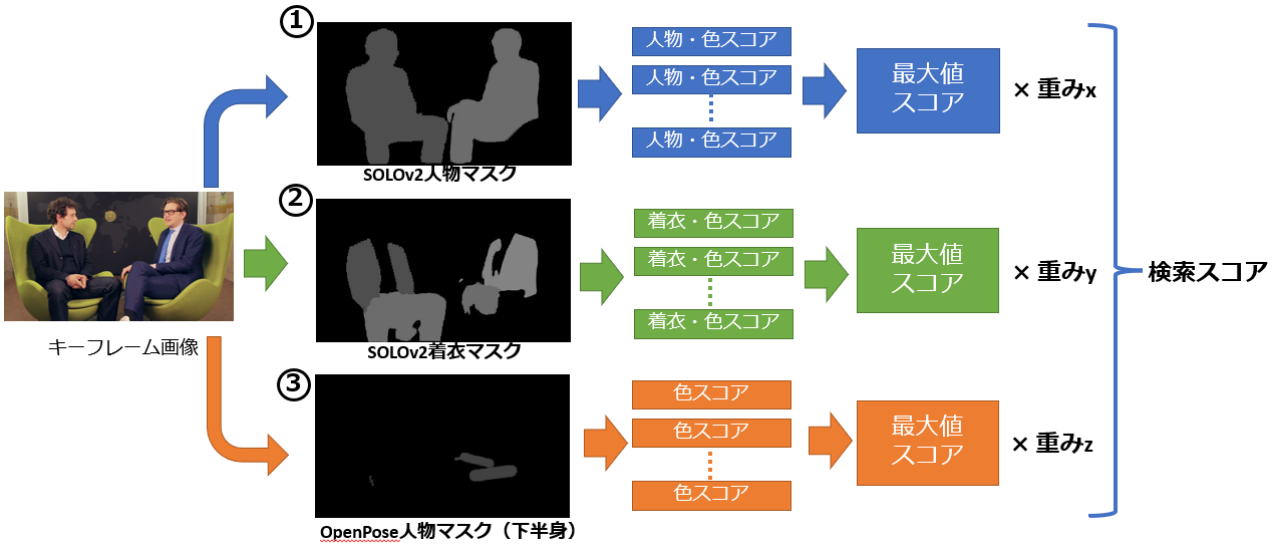

図4　着衣・色識別モジュールの概要図．入力キーフレーム画像から SOLOv2 人物マスク（図中 1），SOLOv2 着衣マスク（図中 2），OpenPose 人物マスク（図中 3）を検出する．この例では，OpenPose 人物マスクは下半身のものを使用している．各手法で検出したマスクそれぞれに対して，マスク領域の色情報と SOLOv2 の分類スコアからスコアを求める．各種法のマスクあたりのスコアの最大値を合計することで入力キーフレーム画像の着衣・色検索スコアとする．また，各マスクのスコアの重要度に応じて，重み x，y，z を設定する．

3.2.1　SOLOv2 人物領域マスクの生成／人物・色スコアの算出

SOLOv2 の人物領域マスクからは，検出された人物領域それぞれの SOLOv2 の分類スコア（人物である確率）と領域の色情報を用いて検索クエリに示された人物・色スコアを求める．

キーフレーム画像から人物領域を検出するための SOLOv2 のモデルは，MSCOCO データセットを ResNet50 のモデルで学習した学習済みモデル[*1]を利用した．マスク画像出力の概要図を図5に示す．入力は 1 つの画像で，出力は人物のマスク画像と，そのマスクに対応した SOLOv2 のカテゴリ分類のスコア（確率）である．人物のマスクは，出力確率と対応付けを行うため，ID 付けをする．ID は，各マスクの SOLOv2 の出力確率のスコアが高い順に 1 から振り分け，マスクを ID の値を画素値としてマスクを出力する．

検出した SOLOv2 人物領域マスクを用いて，マスクごとの色・人物スコアを算出する．元画像に対して，検出したマスクの領域からランダムに 1,024 ピクセルの HSV 色情報の色相の値を取得する．このとき，極端に彩度や明度が低い領域を排除するため，彩度と明度がそれぞれ 25% 以上の領域のみを色情報の抽出の対象とする．また，検索対象の着衣のより近い領域の色情報を取得するため，対象の着衣が上半身のものならマスク領域の上半分，下半身のものならマスク領

[*1] https://cloudstor.aarnet.edu.au/plus/s/DvjgeaPCarKZoVL/

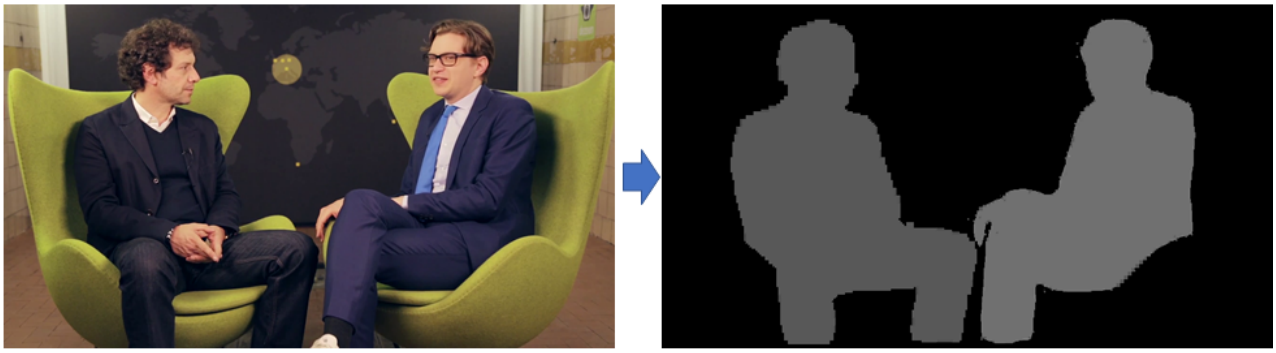

図5　SOLOv2 を用いた人物のマスク検出の例．人物ごとに SOLOv2 の出力確率順に画素値 1，2…でマスクを出力する．この画像では，左の人物を画素値 1，右の人物を画素値 2 で出力している．図の画像は見やすいように明るさ・コントラストの調整を行っている．

域の下半分の領域から色情報を取得する．具体的には，入力された着衣のカテゴリ ID が 0〜5，9〜12 なら上半分，6〜8 なら下半分の領域を使用する．この処理は，人物領域のマスクから検索対象の着衣を着ていると考えらえれる領域の色情報を取得することができる．取得した 1,024 ピクセルのすべての色相の値に，検索する色相の値を中央値とした正規分布に基づいた重みを掛け，合計したものをカラースコアとする．重みを求める正規分布の式を式（1）に示す．

$$f(x) = e\frac{-(x-\mu)^2}{2\sigma^2} \tag{1}$$

このとき，x は 0〜180 の入力された色相の値とする．また，今回の手法では，$\mu = 0$，$\sigma = 10$ とした．例として，$x = 120$（青）としたときの重みの例を図6に示す．

算出したカラースコアと SOLOv2 の分類スコア（人物である確率）の乗算をマスクごとの色・人物スコアとする．これを検出したマスクの数だけ算出し，もっとも大きい色・人物スコアを SOLOv2 人物マスクの出力スコアとする．

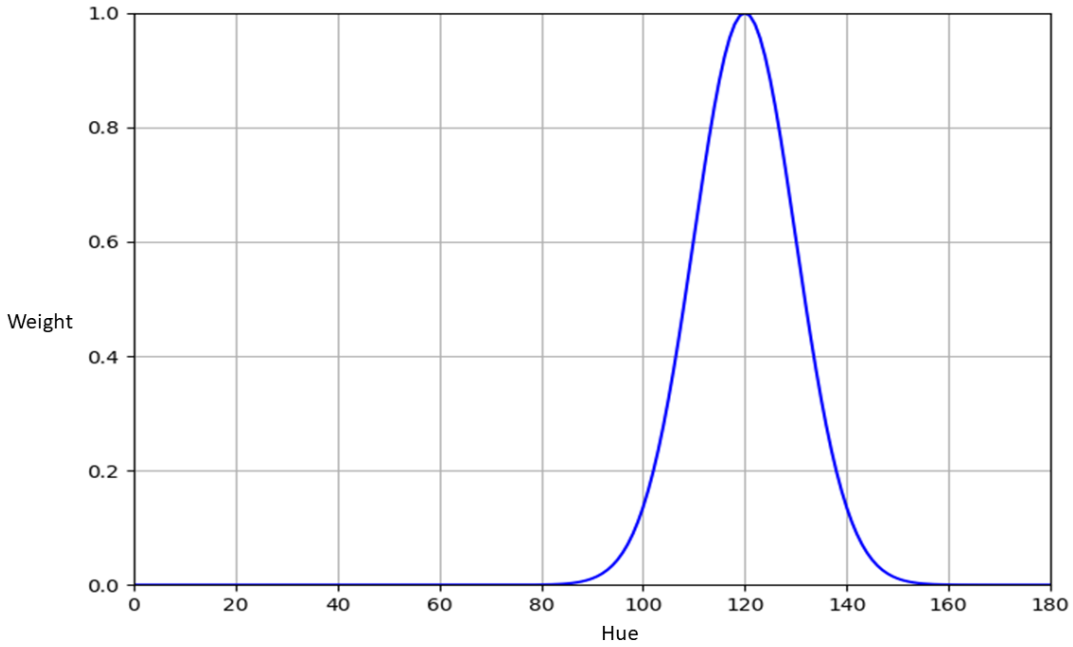

図6　青色（色相 = 120）を中心としたときの重みの例

3.2.2　SOLOv2 を用いた着衣領域マスクの生成／着衣・色スコアの算出

SOLOv2 着衣領域マスクからは，入力キーフレーム画像から検出された着衣領域それぞれの SOLOv2 のカテゴリ ID・分類スコア（そのカテゴリ ID の着衣である確率）と領域の色情報を用いて検索クエリに示された着衣・色スコアを求める．

キーフレーム画像中の着衣領域を検出するために，大規模なファッション画像のデータベースである DeepFashion2 を ResNet18 で学習した SOLOv2 のモデルを使用した．DeepFashion2 は，491,895 個のファッション画像をもつデータベースであり，各画像に 13 の服のタイプを示すクラスラベル，バウンディングボックス，セグメンテーション，ランドマークが与えられている．13 の着衣のタイプの中には，Short sleeve top や vest のような上半身に身につける着衣と，Short，Trousers といった下半身に身につける着衣のカテゴリが含まれる．また，同じ着衣のなかで，店舗が撮影した画像，ユーザーが撮影した画像が与えられている．DeepFashion2 の画像例を図7に示す．また，カテゴリ，ランドマーク，セグメンテーションの例を図8に示す．

入力キーフレーム画像に対して，DeepFashion2 で定められた 13 種のカテゴリの着衣のマスク画像と，カテゴリ ID（0 12），SOLOv2 の分類スコアを出力する．カテゴリ ID，SOLOv2 の分類スコアは，検出された着衣マスクの数だけ出力される．出力マスク画像の例を図9に示す．

検出した SOLOv2 着衣領域マスクを用いて，マスクごとの色・着衣スコアを算出する．カラースコアの抽出方法は，節3.2.1と同様にマスク領域から出力する．抽出したカラースコアと，SOLOv2 の分類スコアの乗算を色・着衣スコアとして出力する．検出したマスクの数だけこの処

図7　DeepFashion2 の画像例

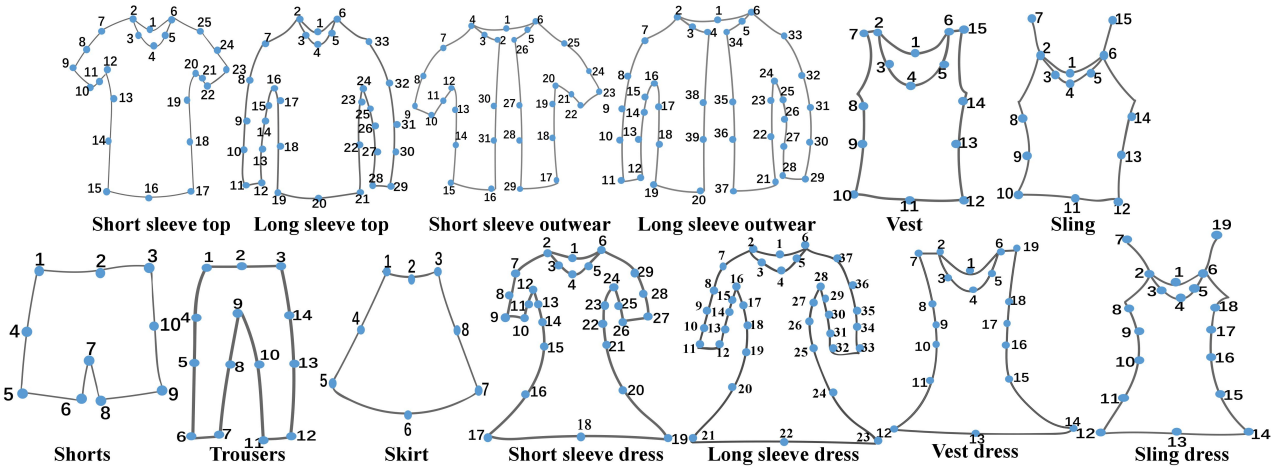

図8　DeepFashion2 カテゴリ，ランドマーク，セグメンテーションのアノテーション一覧
（https://github.com/switchablenorms/DeepFashion2 より引用）

理を行う．このとき，着衣マスクのカテゴリが検索対象となる着衣の ID ではなかった場合，そのマスクの色・着衣スコアは 0 として扱う．すべてのマスクの色・着衣スコアの最大値を着衣領域マスクの出力スコアとする．

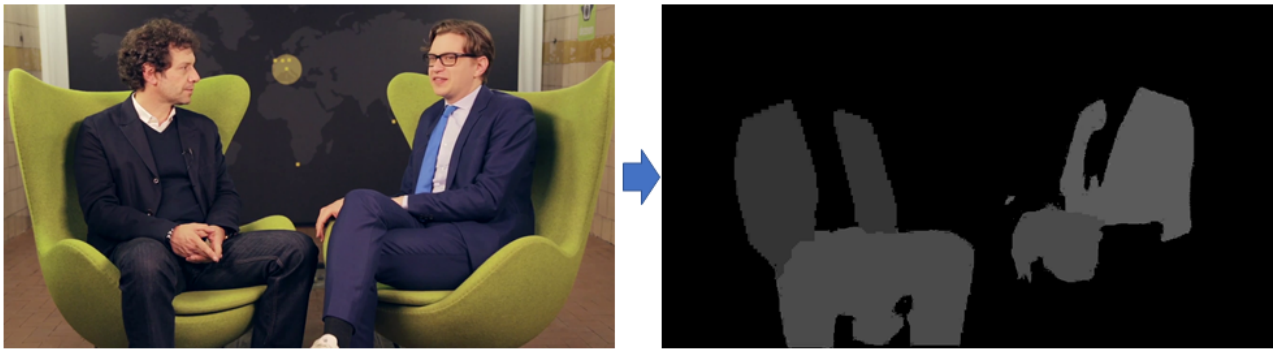

図9　SOLOv2 を用いた人物のマスク検出の例．人物ごとに SOLOv2 の出力確率順に画素値 1，2…でマスクを出力している．例の画像は見やすいように明るさ・コントラストの調整を行っている．

3.2.3　OpenPose を用いた人物領域マスクの生成／色スコアの算出

本モジュールでは，OpenPose から得られる人物の関節点情報を用いて，人物のマスク画像を生成する．

OpenPose は関節点の情報から，体の部位ごとに座標を得ることができる．そのため，検索対象となる着衣によって色情報を抽出する領域を制限することができる．

本手法では，関節点情報を用いることで，上半身，下半身の領域を分割し，マスク画像を生成する．具体的には，右肩・左肩・右腰・左腰を結んだ四角形の領域を上半身領域のマスク，右腰-右ひざ-右足首と左腰-左ひざ-左足首を結んだ線分を下半身領域のマスクとする．上半身／下半身のマスク画像の例を図10に示す．

検出した OpenPose 人物領域マスクを用いて，マスクごとのカラースコアを算出する．カラースコアの抽出方法は，節3.2.1と同様にマスク領域から出力する．色情報を抽出するマスク領域は，入力された着衣のカテゴリが上半身の着衣か下半身の着衣かで設定する．具体的には，入力された着衣のカテゴリ ID が 0～5・9～12 なら上半身領域のマスク，6～8 なら下半身領域のマスクを使用する．検出されたすべてのマスクからカラースコアを算出し，その最大値を OpenPose 人物領域マスクの出力スコアとする．

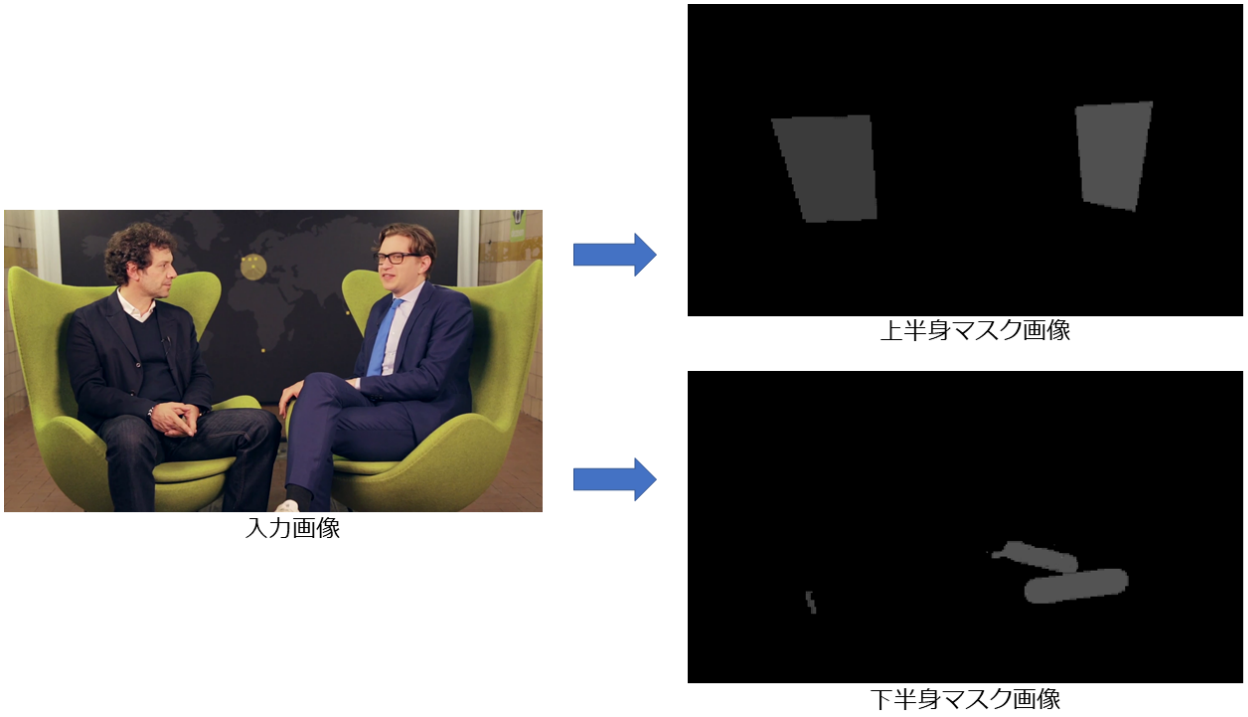

図10　OpenPose を用いた人物のマスク検出の例．関節点の座標から，上半身と下半身のマスクを生成する．検索したい着衣によって，どちらのマスクから色情報を抽出するかを決定する．例の画像は見やすいように明るさ・コントラストの調整を行っている．

3.3　各モジュールのスコアの統合

画像-言語類似度算出モジュールと着衣・色識別モジュールを用いて，キーフレーム画像 1 個ごとの統合検索スコアを出力する．各モジュールのスコアは，min-max 正規化を用いて 0 から 1 の範囲に正規化する．その後，両モジュールのスコアに重み（倍率）W_1，W_2 を設定する．重みを乗算した両モジュールのスコアを合計し，キーフレーム画像 1 個に対する検索スコアとする．この検索スコアをすべてのキーフレーム画像について算出し，ランキングを行うことで映像の検索を行う．

4　大規模映像データベースを用いた検索実験

提案手法の有効性を検証するため，大規模映像データから特定の着衣を着た人物について検索実験を行った．実験に使用する映像データベースは，TRECVID ベンチマークの AVS タスクで 2019〜2021 年に用いられた大規模映像データ V3C1 を用いた．V3C1 は, 平均時間 8 分の 7,475 個の映像データ，映像データをシーンごとに分割した 1,082,657 個の短い映像データとそのキーフレーム（静止画）からなる．今回の実験では，キーフレーム画像を検索実験に使用した．V3C1 のキーフレーム画像の例を図11に示す．

図11　V3C1 のキーフレーム画像の例．7,457 の映像データから，シーンごとに分割された 1,082,657 の映像につき 1 つのキーフレーム画像が存在する

実験では，着衣・色識別モジュールに入力する着衣のカテゴリ（DeepFashion2 のカテゴリ，0〜12），着衣の色相の値（0〜180 の範囲）を検索クエリに応じて設定する．また，着衣・色識別モジュール内の SOLOv2 着衣領域マスクの重みは 5 に設定した．

すべてのキーフレーム画像から統合検索スコアを求め，降順にランキングを行う．評価には，上位 20 枚までに検索対象となるキーフレーム画像が目視で確認できるかの定性的評価と，上位 1000 枚までの正解データとの平均適合率を用いた定量的評価を行う．

4.1　実験条件

4.1.1　言語-画像類似度モジュールに用いるエンコーダのモデル

本実験では，言語-画像類似度モジュールに用いるエンコーダに CLIP [3] のモデルを用いる．CLIP は，インターネット上の膨大な数（約 4 億）の画像-文章の関係性を学習したモデルであり，画像・文章から共通の空間の特徴ベクトルを算出することができる．

CLIP は画像中から一般的な幅広い特徴を捉えることができるが，物体の詳細な属性の推定が難しい課題があげられている．提案手法では，画像中の詳細な着衣の特徴を抽出する着衣・色識別モジュールを組み合わせるため，CLIP の幅広い特徴を捉える性能と詳細な着衣特徴を識別する性能を持つ映像検索が行えることが考えられる．

4.1.2　実験に使用する検索クエリ

入力に使う検索クエリは，TRECVID 2019・2020 の AVS タスクの合計 40 個の検索クエリのうち，映像中の着衣の情報が関係すると考えられる以下のクエリについて検索実験を行った．

- Find shots of a man in blue jeans outdoors
- Find shots of a woman wearing a red dress outside in the daytime
- Find shots of a person wearing shorts outdoors

4.1.3　各モジュールの重みの設定

検索実験で用いる各モジュールの重みを設定する．

提案手法の両モジュールのスコア統合で使用する重み W_1，W_2 は，どちらも 1.0 に設定した．

着衣・色識別モジュール内の各マスクのスコアの重みについては，SOLOv2 人物領域マスクのスコアの重み x を 1.0，SOLOv2 着衣領域マスクのスコアの重み y を 5.0，OpenPose 人物領域マスクのスコアの重み z を 1.0 に設定した．

4.2　検索クエリ「Find shots of a man in blue jeans outdoors」

検索クエリを「Find shots of a man in blue jeans outdoors」としたときの検索結果を評価する．検索クエリの「青いジーンズをはいた男性」に従って，着衣・色識別モジュールの入力は，色相を 120（青），着衣のカテゴリを 7（Trousers）に設定した．

上位 20 位までの画像を図12に示す．上位 20 位までの画像を目視で確認したところ，検索対象となる屋外で青いジーンズをはいた人物が確認できた．しかし，青いジーンズを履いているが屋外の画像ではないものも確認できた．これは，提案手法の着衣・色識別モジュールが，着衣カテゴリとその色以外の要素を考慮していないため，それ以外の要素が間違った画像であっても高い検索スコアを出力している可能性が考えられるためである．

また，青い長ズボンは履いているがジーンズではない画像も確認できた．これは，着衣・色識別モジュールが，入力された検索クエリに含まれる着衣の形状をもとに着衣カテゴリのスコアを出力しているため，ジーンズなどの形状以外の着衣の種類を識別することができないことが原因であると考えられる．

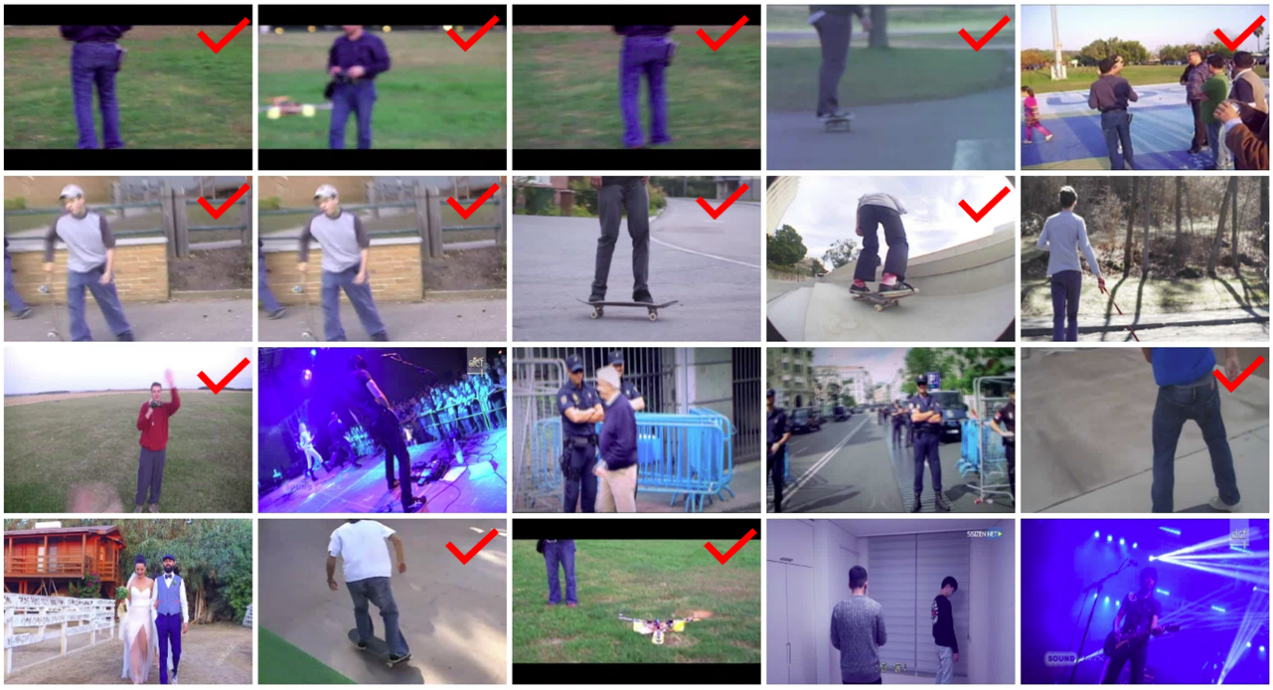

図12　検索クエリを「Find shots of a man in blue jeans outdoors」としたときの検索結果
上位 20 位までの画像．赤いチェックマークが正解画像．

4.3 検索クエリ「Find shots of a woman wearing a red dress outside in the daytime」

検索クエリを「Find shots of a woman wearing a red dress outside in the daytime」としたときの検索結果の評価を行う．「赤いドレスを身につけた人物」に従って，着衣・色識別モジュールの入力は，色相は 0（赤），着衣のカテゴリは 9〜12（Short sleeve dress，Long sleeve dress，Vest dress，Sling dress）に設定した．

上位 20 位までの画像を図13に示す．目視での評価を行ったところ，検索対象となる屋外で赤いドレスを身につけた人物が確認できた．しかし，赤いドレスを身につけていても，屋内に人物がいるような画像も確認できた．これは，着衣・色識別モジュールに背景などの情報を分析する機能がないため，赤いドレスを着ているだけでも高いスコアが算出されている可能性が考えられる．また，上位 1000 枚までの正解データとの AP は 0.2118 であった．

図13　検索クエリを「Find shots of a woman wearing a red dress outside in the daytime」としたときの検索結果上位 20 位までの画像．赤いチェックマークが正解画像．青い × は赤いドレスが画像中に確認できるが，屋内，日中ではないものを示す．

4.4　検索クエリ「Find shots of a person wearing shorts outdoors」

検索クエリを「Find shots of a person wearing shorts outdoors」としたときの検索結果を評価する．着衣・色識別モジュールの入力は，検索クエリの「短いズボンをはいた人物」に従って着衣カテゴリを 6（Shorts）に設定した．また，本検索クエリには着衣の色の指定はないので，着衣・色識別モジュールの各マスクから得られるカラースコアは用いず，SOLOv2 着衣マスクの分類スコア（着衣の確率）のみで検索スコアを算出した．

上位 20 位までの画像を目視で確認したところ，検索クエリに示された通り，「屋外で短いズボンをはいた人物」が確認できた．

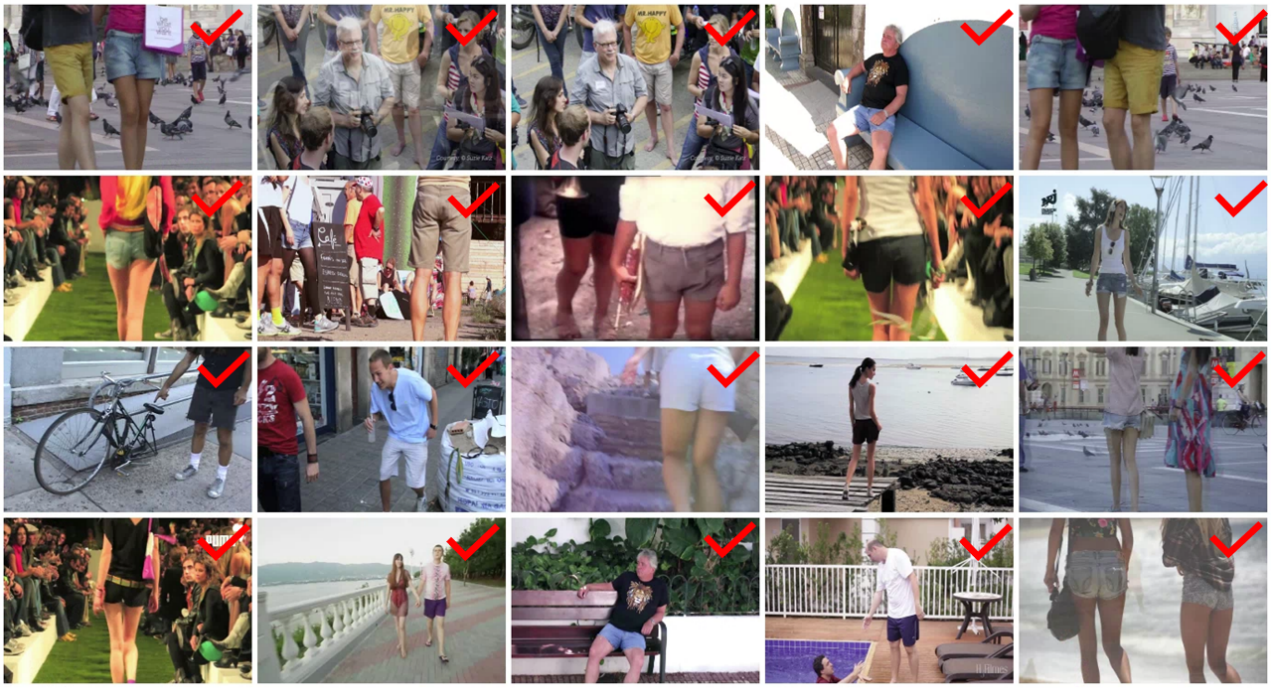

図14　検索クエリを「Find shots of a person wearing shorts outdoors」としたときの検索結果上位 20 位までの画像．赤いチェックマークが正解画像．

5　各モジュールの有効性の検証実験

提案手法では，言語-画像類似度算出モジュールから得られたクエリ-画像類似度スコアと，着衣・色識別モジュールから得られた詳細な着衣の特徴を考慮した着衣・色検索スコアを統合して画像検索を行った．しかし，着衣・色識別モジュールから出力される着衣・色検索スコアは，背景，人物の性別といった着衣以外の要素を考慮していない．そのため，検索の条件から外れている映像についても高いスコアが出力されている可能性があり，言語-画像類似度算出モジュールのクエリ-画像類似度スコアとの統合を行った場合に検索精度に悪影響を与えていることが考えられる．そこで，着衣・色検索スコアに重み（倍率）を設定し，スコアの統合実験を行う．

提案手法の各モジュールの有効性を確認するため，提案手法の言語-画像類似度算出モジュール，着衣・色識別モジュールそれぞれ単体で検索実験を行った場合の結果の検証を行う．言語-画像類似度算出モジュールのみで検索実験を行うときは提案手法の重み W_1 を 1.0，W_2 を 0 に設定する．同様に，着衣・色識別モジュールのみで検索実験を行うときはは提案手法の重み W_1 を 0，W_2 を 1.0 に設定する．

検索対象とするクエリは，4.1.2節に示したものと同じものを使用した．

5.1　検索クエリ「Find shots of a man in blue jeans outdoors」

検索クエリを「Find shots of a man in blue jeans outdoors」にしたときに，言語-画像類似度算出モジュール・着衣・色識別モジュールそれぞれ単体で検索を行った時の結果を評価する．

着衣・色識別モジュールのスコアのみで検索を行った場合の検索結果上位 20 位までの画像を図15に示す．検索結果上位の画像を目視で評価したところ，検索対象に設定した「青い長ズボン」が確認できた．しかし，元となる検索クエリの「青いジーンズ」であるかはわからない画像も存在した．また，検索クエリでは「男性」が指定されているが，長ズボンをはいた女性の画像も確認できる．これは，提案手法では性別を解析する機能は組み込んでいないためである．また，上位 1000 枚までの正解データとの平均適合率は 0.0032 であった．これは，提案手法では検索クエリの着衣以外要素については識別できていないことが原因であると考えられる．元となる検索クエリ「Find shots of a man in blue jeans outdoors」から，検索に必要となる情報は「屋外で」「男性が」「青いジーンズをはいている」の 3 つである．しかし，提案手法では着衣の検出・カテゴリ・色の識別を目的にしているため，「屋外で」「男性が」の 2 つの要素については解析できていない．また，提案手法では「青い長ズボン」を検索対象としているが，検索クエリで示された「青いジーンズをはいている」要素についても正しく解析できていない可能性がある．例えば，「青い長ズボン」のなかでも，検索クエリに示された「ジーンズ」ではなく，スラックスやチノパンツである可能性もあるためである．

また，4.2節の検索結果と同様に，画像全体の色調が青に偏っている画像も確認できた．

表1　検索クエリを「Find shots of a man in blue jeans outdoors」としたとき，各モジュールのみのスコアで検索した結果の平均適合率

手法	平均適合率
提案手法（両モジュール使用）	0.0278
着衣・色識別モジュールのみ	0.0032
言語-画像類似度算出モジュールのみ	0.0568

図15　検索クエリを「Find shots of a man in blue jeans outdoors」としたときの，着衣・色識別モジュールのスコアのみで検索した場合の検索結果上位 20 位までの画像．赤い　が正解画像．青い × は，青いズボンをはいた人物は確認できるが，ジーンズではない，男性ではない，屋外ではないといった，検索クエリに一致しない画像を示す．

着衣・色識別モジュールのスコアのみで検索を行った場合の検索結果上位 20 位までの画像を図15に示す．検索結果上位を目視で評価したとこと，検索対象となる青いジーンズが含まれる画像が多く確認できた．しかし，ジーンズが画像全体に大きく写っている画像では，性別が確認できないような画像も確認できた．また，上位 1000 位までの正解データとの平均適合率は 0.0568 となった．このことから，提案手法でスコア統合を行うとき，着衣以外の要素を解析できない着衣・色識別モジュールのスコアが大きく影響してしまうと検索精度が低下してしまうことがあるとわかった．

図16　検索クエリを「Find shots of a man in blue jeans outdoors」としたときの，言語-画像類似度算出モジュールのスコアのみで検索した場合の検索結果上位 20 位までの画像．赤いチェックマークが正解画像．青い × は，ジーンズをはいているが男性と判断できない（正解画像であると判別できない）ものを示す．

5.1.1　検索クエリ「Find shots of a woman wearing a red dress outside in the daytime」

検索クエリを「Find shots of a woman wearing a red dress outside in the daytime」にしたときに，言語-画像類似度算出モジュール・着衣・色識別モジュールそれぞれ単体で検索を行った時の結果を評価する．

着衣・色識別モジュールのスコアのみで検索を行った検索結果上位 20 位までの画像を図17に示す．図17から，検索対象に設定した赤い Short sleeve dress，Long sleeve dress，Vest dress，Sling dress の着衣が確認できる．しかし，検索クエリの要素である「屋外」「日中」に該当しない画像も確認できた．上位 1000 枚までの正解データとの平均適合率は 0.0875 であった．これは，検索クエリに示される「屋外」「日中」といった要素が提案手法では解析できていないことが原因であると考えられる．また，上位 20 位までの画像（図17）に，ドレスよりは丈が短いワンピースのような着衣が確認できる．これらの着衣は正解データの「赤いドレス」には該当しない可能性が考えられる．

言語-画像類似度算出モジュールのスコアのみで検索を行った検索結果上位 20 位までの画像を図18に示す．目視で評価を行ったところ，検索対象となる屋外，日中に該当する画像が確認でき

た．しかし，着衣がドレスのようには見えない画像も確認できる．また，上位 1000 位までの AP は 0.0525 と着衣・色識別モジュールのみで検索を行った場合より低くなった．提案手法の平均適合率は 0.2118 であることから，言語-画像類似度算出モジュールが苦手とするドレスの着衣の検出の補助に，着衣・色識別モジュールが有効に働いていることがわかる．着衣・色識別モジュールと言語-画像類似度算出モジュールの機能を無効化した場合の検索クエリ「Find shots of a woman wearing a red dress outside in the daytime」について検索した結果の上位 1000 位までの平均適合率を表2に示す．

表2　検索クエリを「Find shots of a woman wearing a red dress outside in the daytime」としたとき，各モジュールのみのスコアで検索した結果の平均適合率

手法	平均適合率
提案手法（両モジュール使用）	**0.2118**
着衣・色識別モジュールのみ	0.0875
言語-画像類似度算出モジュールのみ	0.0525

図17　検索クエリを「Find shots of a woman wearing a red dress outside in the daytime」としたときの，着衣・色識別モジュールのスコアのみで検索した場合の検索結果上位 20 位までの画像．赤いチェックマークが正解画像．青い × は，赤いドレスを身につけてはいるが屋外，日中ではないといった検索クエリに一致しない画像である．

図18　検索クエリを「Find shots of a woman wearing a red dress outside in the daytime」としたときの，言語-画像類似度算出モジュールのスコアのみで検索した場合の検索結果上位20 位までの画像．赤いチェックマークが正解画像を示す．

5.1.2　検索クエリ「Find shots of a person wearing shorts outdoors」

検索クエリを「Find shots of a person wearing shorts outdoors」にしたときに，言語-画像類似度算出モジュール・着衣・色識別モジュールそれぞれ単体で検索を行った時の結果を評価する．

着衣・色識別モジュールのみで検索クエリ「Find shots of a person wearing shorts outdoors」について検索を行った場合の検索結果上位 20 位までの画像を図19に示す．上位 20 位までの画像を目視で確認したところ，検索対象に設定した短いズボンをはいた人物が確認できた．しかし，検索クエリに示されている「屋外」以外の画像も確認できた．これは，着衣・色識別モジュールのみでは着衣以外の情報が得られないためである．また，上位 1000 枚までの平均適合率は 0.1350 であった．これは，本クエリで指定されている「屋外」の要素が解析できていないためであると考えられる．

言語-画像類似度算出モジュールのみで検索クエリ「Find shots of a person wearing shorts outdoors」について検索を行った場合の検索結果上位 20 位までの画像を図20に示す．上位 20 位までの画像を目視で確認したとこと，検索クエリに含まれる短いズボンをはいた人物が確認できる．しかし，スカートのように見える着衣を身につけた人物や，画像全体に足の肌が写っていて着衣を判別できないような画像も確認できた．言語-画像類似度算出モジュールのみで映像検索

を行った際の上位 1000 位までの平均適合率は 0.0886 であった．このことから，着衣・色識別モジュールと言語-画像類似度算出モジュールのスコアを統合することで，CLIP の特徴抽出の補助に有効であることがわかった．着衣・色識別モジュールと言語-画像類似度算出モジュールの機能を無効化した場合の検索クエリ Find shots of a person wearing shorts outdoors」について検索した結果の上位 1000 位までの AP を表3に示す．

表3　検索クエリを「Find shots of a person wearing shorts outdoors」としたとき，各モジュールのみのスコアで検索した結果の平均適合率

手法	平均適合率
提案手法（両モジュール使用）	**0.2360**
着衣・色識別モジュールのみ	0.1350
言語-画像類似度算出モジュールのみ	0.0886

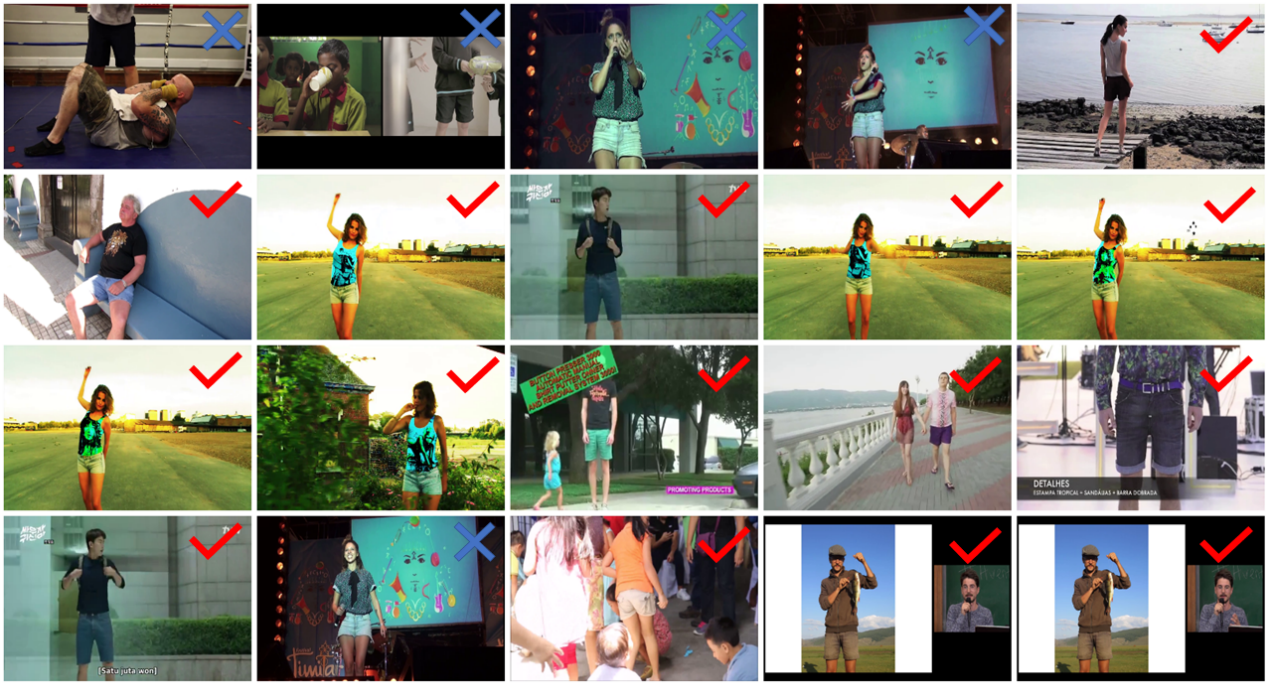

図19　検索クエリを「Find shots of a person wearing shorts outdoors」としたときの，着衣・色識別モジュールのスコアのみで検索した場合の検索結果上位 20 位までの画像．赤いチェックマークが正解画像を示す．青い × は，短いズボンをはいた人物が確認できるが，検索クエリに示された屋外ではない画像を示す．

図20　検索クエリを「Find shots of a person wearing shorts outdoors」としたときの，言語-画像類似度算出モジュールのスコアのみで検索した場合の検索結果上位 20 位までの画像．赤いチェックマークが正解画像を示す．

5.2　各モジュールの有効性の検証：考察

着衣・色識別モジュールのみで映像検索を行った場合，いずれの検索クエリでも検索対象として指定した着衣が上位の画像に確認できた．しかし，提案手法は人物・着衣領域以外の情報を解析する機能を組み込んでいないため，背景や着衣以外の人物の属性が関係する検索クエリだと正しく検索できていない問題があった．映像検索の際に入力される検索クエリは人物の属性から背景など幅広い意味を持つため，検索精度向上にためには画像から着衣以外の情報を解析する機能が必要である．

言語-画像類似度算出モジュールのみで映像検索を行った場合，上位の画像に検索クエリの対象となると考えられる画像が確認できた．また，屋外，時間帯などの着衣以外の要素についても特徴をとらえた映像検索ができていた．しかし，検索クエリに示された着衣とは別の着衣を身につけた人物や，着衣そのものが画像全体の大きな領域を占めていて着衣以外の要素が確認できないような画像も確認できた．

また，上位 1000 位までの平均適合率から，着衣・色識別モジュールと言語-画像類似度算出モジュールを組み合わせることで，検索精度の向上に貢献している検索クエリも存在した．しかし，片方のモジュールで誤検出が増加した場合，検索精度が低下したような検索クエリもあった．

6　スコア統合時の重みを変化させたときの検索結果の分析

提案手法の着衣・色識別モジュールから出力される着衣・色検索スコアは，背景，人物の性別といった着衣以外の要素を考慮していない．そのため，検索の条件から外れている映像についても高いスコアが出力され，検索精度の低下につながっている可能性がある．そこで，着衣・色識別モジュールの検索スコアの重み W_2 を変化させ，検索結果への影響を調査した．

着衣・色検索スコアに与える重み W_2 は，1.0，0.5，0.1，0.05 とした．各重みを適用してスコア統合を行い，検索実験を行う．評価には，上位 1000 位までの正解データとの平均適合率を用いる．各重みに対する結果（平均適合率）の一覧を表4に示す．

どの検索クエリにおいても，重みを 1 から 0.5 に変更することにより精度の向上が見られた．言語-画像類似度算出モジュールの大域的な視覚的特徴を考慮したスコアに対して，着衣・色識別モジュールのスコアを補助的に統合することが有効であると考えられる．また，検索クエリ「Find shots of a woman wearing a red dress outside in the daytime」「Find shots of a person wearing shorts outdoors」においては，着衣・色識別モジュールのスコアの重みを 0.1 に設定したとき，重み 1.0 の場合より精度が低下した．このことから，着衣・色識別モジュールのスコアが検索精度の向上に貢献していることがわかる．

表4　着衣・色識別モジュールのスコアに重みを適用したときの上位 1000 位までの平均適合率．検索クエリの blue jeans は「Find shots of a man in blue jeans outdoors」，red dress は「Find shots of a woman wearing a red dress outside in the daytime」，shorts は「Find shots of a person wearing shorts outdoors」に対応している．

	検索クエリ		
着衣スコアの重み	blue jeans	red dress	shorts
1	0.0278	0.2118	0.2360
0.5	0.0797	**0.2572**	**0.2477**
0.1	**0.0824**	0.1227	0.1813
0.05	0.0637	0.0841	0.1269

7　まとめ

本論文では，言語と画像の類似度を算出できる言語-画像埋め込み技術をベースに，映像中の着衣とその色を詳細に識別する機能を組み合わせ，大規模映像データベースから特定の種類の着衣とその色を検索するシステムを提案した．

提案したシステムは，CLIP を用いて検索クエリと画像の類似度を算出する言語-画像類似度算出モジュールと，映像中の人物・着衣領域を検出して着衣の識別・色情報の抽出を行う着衣・色識別モジュールからなる．

言語-画像類似度算出モジュールでは，入力された検索クエリと映像の特徴量の類似度を計算し，検索クエリの条件を満たす映像データを検索する．

着衣・色識別モジュールでは，SOLOv2・姿勢推定で着衣領域の検出と着衣の識別を行い，領域から色情報を抽出することで検索クエリで指定された特定の着衣をとその色を検索する．

2 つのモジュールから得られたスコアを合計することで，幅広い意味をもつ検索クエリの特徴を含みつつ，特定の着衣を身につけた人物が写った映像を検索する．

提案したシステムを用いて特定の着衣を検索対象に含む検索クエリを用いて映像検索を行ったところ，検索結果上位の映像に検索クエリに該当する映像が確認できた．しかし，着衣・色識別モジュールに着衣以外の要素を推定する機能がないため，着衣以外の要素が検索クエリを満たさない画像も確認できた．

提案したシステムの各モジュールの有効性を検証するため，モジュール単体で検索実験を行った．着衣・色識別モジュールのみで特定の着衣を検索対象に含む検索クエリを用いて映像検索を行ったところ，検索結果上位の画像には検索対象として設定した着衣が確認できた．しかし，着衣以外の人物の属性や，シーンの情報などを解析する機能を組み込んでいないため，着衣以外の要素について正しく検索できていなかった．また，言語-画像類似度算出モジュールのみで検索を行った場合，着衣以外の要素も識別できていたが，検索クエリで指定された着衣が検索できていないケースも確認できた．このことから，両モジュールを組み合わせることで，検索クエリの特徴をとらえつつ着衣・色識別モジュールによる詳細な着衣検索機能をもった映像検索が実現できていることがわかった．しかし，着衣・色識別モジュールの出力スコアが検索結果に悪影響を与えているケースも確認できた．そこで，着衣・色識別モジュールの出力スコアに重みを設定して映像検索を行ったところ，重みを設定しない場合より検索精度の向上が見られた．

8　今後の課題

今後の課題として，提案手法の着衣・色識別モジュールへ入力する着衣カテゴリ ID は，入力検索クエリから読み取れる着衣の形状をもとに設定している．しかし，現在の手法では着衣の形状以外の情報を考慮していないため，検索クエリによっては検索に必要な着衣の属性を得ることができない．例として，検索クエリに「ジーンズ」という着衣情報が入っていても，着衣・色識別モジュールは「長ズボン」であるかどうかしか識別できないため，ジーンズではない長ズボンの画像が上位に検索されることがある．そこで，形状以外の着衣のカテゴリを判別する機能を追加することで，より詳細な着衣の属性に着目した映像検索を行うことを予定している．

また，本論文で提案した手法は，あらかじめ検索対象となる大規模映像データベースのすべてのキーフレーム画像から各モジュールの特徴ベクトルを出力しておき，入力された検索クエリの特徴ベクトルとの類似度から映像検索に用いるスコアを算出するシステムとなっている．このシステムの実運用を考慮すると，検索する映像データベースが変化するたびにすべての映像のキーフレーム画像の特徴ベクトルを出力する必要があり，計算コストの増加や検索速度の低下が予想される．この問題を解決するため，各モジュールに入力するキーフレーム画像を入力前の段階で検索クエリに沿って選別するようなシステムの導入を考えている．

謝辞

本研究にあたり，研究方針，プログラムの記述法など，長期にわたりご指導してくださった植木一也推教授に，心より感謝申し上げます．また，研究や技術について日々議論を交えていただいた植木研究室の皆様に深く感謝いたします．

発表実績

1. 堀達史，武藤良，植木一也，
   “多段階の敵対的生成ネットワークによる人物除去,”
   第 26 回画像センシングシンポジウム (SSII2020), 2020.

2. セイエドネシャドロスタム，武藤良，植木一也，
   “OpenPose を用いた特定の色の服を着た人物の検出,”
   第 26 回画像センシングシンポジウム (SSII2020), 2020.

3. Kazuya Ueki, Tomoka Kojima, Ryou Mutou, Rostam Sayyed Nezhad, Yasuaki Hagiwara
   “Recognition of Japanese Connected Cursive Characters Using Multiple Softmax Outputs,”
   Proceedings of the IEEE 3rd International Conference on Multimedia Information Processing and Retrieval (MIPR2020), 2020.

4. 武藤良，セイエドネシャドロスタム，植木一也，堀隆之，金容範，鈴木裕真，
   “学習済みモデルを用いた大規模映像データにおける特定の色の着衣をつけた人物の検索,”
   ビジョン技術の実利用ワークショップ (ViEW2020), 2020.

5. Kazuya Ueki, Ryo Mutou, Takayuki Hori, Yongbeom Kim, Yuma Suzuki,
   “Zero-shot video retrieval using concept-based and visual-semantic embedding approaches,”TRECVID 2020 Workshop, 2020.

6. Kazuya Ueki, Ryo Mutou, Takayuki Hori, Yongbeom Kim, Yuma Suzuki,
   “Waseda_Meisei_SoftBank at TRECVID 2020: Ad-hoc Video Search,”
   Notebook paper of the TRECVID 2020 Workshop, 2020.

7. 山本啓斗，武藤良，植木一也，堀隆之，金容範，鈴木裕真，
   “少数画像をもとにした顔属性データセットの拡張,”
   動的画像処理実用化ワークショップ (DIA2021)，2021.

8. セイエドネシャドロスタム，武藤良，植木一也，堀隆之，金容範，鈴木裕真，
“服装の色を用いた人物検索に向けた学習済みモデルの活用,”
動的画像処理実用化ワークショップ (DIA2021)，2021.

9. 武藤良，植木一也，堀隆之，金容範，鈴木裕真，
“映像検索に向けた SOLOv2 による着衣の検出と識別,”
第 27 回画像センシングシンポジウム (SSII2021), 2021.

10. 武藤良，植木一也，堀隆之，金容範，鈴木裕真，
“着衣のタイプ・色をもとにした大規模映像からの人物の検索,”
第 24 回画像の認識・理解シンポジウム (MIRU2021), 2021.

11. 千葉晃裕，鈴木和也，小湊勇弥，武藤良，植木一也
“非接触型入力デバイスを用いた人工知能の体験を目的としたゲームの開発,”
NICOGRAPH 2021，2021.

12. Ryou Mutou, Kazuya Ueki, Takayuki Hori, Yongbeom Kim, Yuma Suzuki,
“Human Retrieval from Large-Scale Video Data Based on Types and Colors of Clothing,”
Proceedings of the International Workshop on Advanced Image Technology (IWAIT2022), 2022.

参考文献

[1] National Institute of Standards and Technology. 'TREC Video Retrieval Evaluation: TRECVID'. 27-Jan-2020 https://trecvid.nist.gov/

[2] Xiujun L, Xi Y, Chunyuan L, Pengchuan Z, Xiaowei H, Lei Z, Lijuan W, Houdong H, Li D, Furu W, Yejin C, Jianfeng G, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," arXiv:2004.06165

[3] Alec R, Jong Wook K, Chris H, Aditya R, Gabriel G, Sandhini A, Girish S, Amanda A, Pamela M, Jack C, Gretchen K, Ilya S, "Learning Transferable Visual Models From Natural Language Supervision" , 2021.

[4] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN," In Proc. of the International Conference on Computer Vision (ICCV), 2017.

[5] D. Bolya, C. Zhou, F. Xiao, Y. J. Lee, "YOLACT: Real-time Instance Segmentation," In Proc. of International Conference on Computer Vision (ICCV), 2019.

[6] X. Wang, T. Kong, C. Shen, Y. Jiang, L. Li, "SOLO: Segmenting Objects by Locations," In Proc. of Europian Conference on Computer Vision (ECCV) 2020.

[7] X. Wang, R. Zhang, T. Kong, L. Li, C. Shen, "SOLOv2: Dynamic and Fast Instance Segmentation, In Proc. of Advances in Neural Information Processing Systems (NeurIPS'20), 2020.

[8] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," arXiv:1812.08008, 2018.

[9] R. A. Güler, N. Neverova, I. Kokkinos, "DensePose: Dense Human Pose Estimation In The Wild," In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[10] Ning Z, Manohar P, Marc'Aurelio R, Trevor D, Lubomir B, "PANDA: Pose Aligned Networks for Deep Attribute Modeling," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1637-1644

[11] Z Liu, P Luo, S Qiu, X Wang, X Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations" IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1096-1104

[12] W Wang, Y Xu, J Shen, S Zhu, "DAttentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification," In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[13] Y. Ge, R. Zhang, X. Wang, X. Tang, P. Luo, "DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images," In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.