論文

エッジ空間パターンを利用したテレビ番組映像からの字幕テキスト領域検出

A Localization Method of Video Text in TV Program Video using Edge Space Pattern

正会員 河合 吉彦[†], 正会員 望月 貴裕[†], 正会員 住吉 英樹[†]Yoshihiko Kawai[†], Takahiro Mochizuki[†] and Hideki Sumiyoshi[†]

あらまし 本論文では、エッジ空間パターンを利用したテレビ番組映像からの字幕テキスト領域の検出手法を提案する。エッジ空間パターンは、文字を形成する線や点の配置を反映した特徴量であり、着目画素の周囲におけるエッジの分布パターンに基づいて算出される。エッジの交差や直線以外のエッジも考慮することができ、背景画像の変動にも頑健な特徴量である。提案手法では、番組映像から取得したフレーム画像を走査窓で走査し、窓領域から算出した特徴量を機械学習で判別することによって字幕テキストの候補領域を求める。その後、検出された候補領域をエッジ密度や領域の形状などに基づいて選別し、字幕テキスト領域の外接矩形を検出する。約 10 時間のテレビ番組映像を対象とした評価実験では、再現率が 89.9%、適合率が 88.0%、F 値が 0.889 という結果が得られ、従来手法よりも F 値が 0.135 向上することが確認された。

キーワード：字幕テキスト，字幕スーパー，オープンキャプション，エッジ空間パターン，テレビ番組映像，映像インデキシング

1. ま え が き

近年、放送の多チャンネル化や録画機器の大容量化により、個人でも数千時間規模のテレビ番組映像を保持することが可能となっている。蓄積された大量の映像データを有効活用するには、目的のシーンを簡単に見つけ出すための技術が重要となる。そのためには、そのシーンに何が映っているのか、あるいはそのシーンの主題は何か、といった意味内容の解析が必要不可欠である。意味内容の解析手法には様々なアプローチがあるが^{1)~6)}、その一つに映像に重畳される字幕テキスト（本稿では字幕スーパー、オープンキャプションの意味で用いる）を解析するものがある^{7)~14)16)}。字幕テキストは、被写体の名称や属性、撮影場所、発言の要旨、取り上げている話題、スポーツにおけるスコアといった意味的に高次の情報を含んでおり、映像意味内容解析の有効な手がかりとなる。また、字幕テキストの解析結果はテキストデータとして獲得できるため、キーワードベースの検索システムと親和性が高いという特徴もある⁷⁾。しかしながら、複雑な背景に重畳される字幕テキストを精度よく検出することは非常に難しく、未だ解決困難な課題の一つとなっている⁵⁾⁷⁾。字幕テキストを解析する手順は、おおまかに以下ようになる。

1. 字幕テキストの検出

2. 字幕テキストの認識

字幕テキストの検出処理は、映像に字幕テキストが表示されているかどうかを判定し、フレーム画像内における表示位置を特定する処理である。また、字幕テキストの認識処理は、OCR (Optical Character Recognition) 技術を利用して、字幕テキストに書かれた文字列画像をテキストデータに変換する処理である。本研究では、字幕テキストの検出までを対象とする。具体的には、映像から取得したフレーム画像を入力とし、字幕テキストの外接矩形を検出する。処理対象はテレビ番組映像とする。なお、字幕テキストの認識については、既存の OCR 技術の利用を想定する。

字幕テキストの検出については、従来から様々な研究がなされている。例えば、複数フレームに渡って同じ位置に出現するエッジ成分を抽出し、それらが密集している領域を字幕テキストとして検出する手法がある^{8)~10)}。この手法は、字幕テキストは一定の時間、同じ位置に表示され続けるという仮定に基づいている。表示位置や文字サイズが時間的に変化する場合は検出できない。それに対し、エッジの動きを推定することによって移動する字幕テキストに対応した手法が提案されている¹¹⁾。また、フレーム間の関係を利用せず、単一のフレーム画像から検出できる手法もある^{2)12)~15)}。しかし、いずれの手法も経験則に基づく知識に強く依存しており、様々なジャンルの映像から頑健に検出することが難しいという問題がある。機械学習を利用して

2014 年 11 月 7 日受付, 2015 年 2 月 26 日再受付, 2015 年 4 月 8 日採録

[†] NHK 放送技術研究所

(〒 157-8510 世田谷区砧 1-10-11, TEL 03-5494-3387)

精度向上を図った手法¹⁶⁾もあるが、計算コストや検出精度の面で課題が残されている。

そこで本論文では、テレビ番組映像を対象に、走査窓と機械学習を利用した字幕テキスト領域の検出手法を提案する。あわせて、字幕テキストとそれ以外の領域を精度よく判別するため、エッジ空間パターンと呼ぶ新たな画像特徴量を提案する。本手法では、フレーム画像全体を走査窓で走査し、窓領域から求めた画像特徴量を機械学習で判別することによって字幕テキストの矩形領域を検出する。機械学習にはサポートベクターマシン (SVM: Support Vector Machine) を利用する。実験では、実際の放送番組映像を利用した比較実験により、提案手法の有効性を検証する。

本論文の構成は以下のとおりである。第2章では、映像からのテキスト検出に関わる従来研究を調査し、本研究の位置づけを明確にする。第3章では、提案手法で利用する画像特徴量や検出手順について詳細を説明する。第4章では、テレビ番組映像を利用した評価実験により、提案手法の有効性を検証する。最後に、第5章でまとめと今後の課題について述べる。

2. 関 連 研 究

映像に出現するテキスト情報は、大きく次の2種類に分類できる。一つは映像に重畳される字幕テキストであり、もう一つは情景内の看板や標識などに含まれる文字列 (シーンテキストと呼ぶ) である。

まず、字幕テキストを対象とした従来研究について述べる。鷲尾ら⁸⁾は、複数フレームに渡って同じ位置に現れるエッジ成分と、色相、彩度、輝度が変化しない画素を抽出し、それらが密集している領域を字幕テキストとして検出した。また佐藤ら⁹⁾は、MPEG映像を対象として、高周波成分を多く含み、かつ静止しているマクロブロックを求めることで字幕テキスト領域を検出した。これらは、字幕テキストの表示位置が時間的に変化しないことを前提とした手法といえる。新井ら¹¹⁾は、移動する字幕テキストに対応するため、1次元ハフ変換を用いてフレーム間のエッジの移動を推定する手法を提案した。字幕テキストの判定には、エッジの密集度合や、時間的な継続性に基づく経験的規則を利用した。またZhongら¹⁴⁾は、DCT (Discrete Cosine Transform) 圧縮された単一のフレーム画像から、DCT係数の分布に基づいて字幕テキスト領域を検出した。フレーム間の相関を利用せず、単一フレームから検出できるようにすることで、位置や文字サイズが変化する字幕テキストに対応した。Lienhartら¹⁶⁾は、検出精度の向上を図るため、経験則ではなくニューラルネットワークを利用した判定手法を提案した。

次に、シーンテキストに対応した手法について述べる。情景内に映るシーンテキストは、フレーム画像内での表示位置や向きなどが時間的に変動するため、ほとんどの従来手法は単一のフレーム画像から検出する方式となっている。

Zhangら¹⁷⁾は、エッジ画像から求めたHOG (Histogram of Oriented Gradients) 特徴⁴⁾を用いて候補領域を求め、グラフスペクトラムで統合する手法を提案した。またBaiら¹⁸⁾は、Harrisコーナー検出器を利用して候補位置を求め、色や位置の類似性に基づいてテキスト領域の外接矩形を検出した。Epshteinら¹⁹⁾は、Stroke Width Transform (SWT) と呼ばれる手法を提案した。SWTでは、エッジの勾配方向に基づいて局所的なストローク幅を求め、同様のストローク幅を持つ画素をグループ化することでテキスト領域を検出した。また、機械学習に基づく手法として、Wangら²⁰⁾は小領域ごとに輝度のコントラストやエッジ方向ヒストグラムなどを求め、SVMでテキスト領域か否かを判定した。Minettoら²¹⁾は、SnooperTextと呼ばれるアルゴリズムを提案している。まず、フーリエ変換やゼルニケ近似多項式に基づく特徴量をSVMで判別して文字領域を検出した後、近接した文字領域をグループ化することでテキスト候補領域を求めた。さらに後処理として、テキスト候補領域から算出したT-HOG (HOG-based Texture) 特徴²²⁾をSVMで判定し、最終的な検出結果とした。その他にも、ニューラルネットワークを利用した手法²³⁾や、Real AdaBoost²⁴⁾を利用した手法²⁵⁾などが提案されている。シーンテキスト検出手法の多くは、字幕テキスト検出手法を発展させ、カメラの傾きなどに対応する処理を追加したものとなっている。そのため、これらの手法は字幕テキストの検出にも適用できる場合がある。

提案手法は、Wangら²⁰⁾やMinettoら²¹⁾の手法と同様のアプローチといえる。つまり、小領域から求めた画像特徴量を機械学習で判別することによって、テキスト候補領域を求めるような方法である。提案手法の特徴の一つは、小領域からの特徴量の算出方法である。提案手法では、小領域における文字らしさを判定するため、エッジ空間パターンと呼ぶ新たな特徴量を提案する。エッジ空間パターンは、文字を形成する線や点の配置を反映した特徴量である。エッジの交差や直線以外のエッジも考慮でき、背景画像の変動にも頑健な特徴量である。提案手法では、エッジ空間パターンを含む複数の画像特徴量を組合せることにより、字幕テキストとそれ以外の領域を精度よく判別する。もう一つの特徴は、テキスト候補領域の検出方法である。従来手法では、始めに文字領域を検出し、文字単位でテキスト領域か否かを判定していたが²⁰⁾²¹⁾、提案手法では、フレーム画像全体を走査窓で走査してテキスト候補領域を検出する。そのため、従来手法とは異なり、走査窓が文字の境界を跨いだ場合でも候補領域と判定できるよう考慮している。具体的には、小領域内の文字の位置や数に依存しない画像特徴量を利用し、機械学習には文字の境界を無視して無作為に切り出した画像を利用する。これにより、文字単位への誤分割に起因する検出精度の低下を避けるとともに、文字境界の検出に要する計算コストの削減を図る。

図 1 提案手法の概要

3. 提案手法

図 1 に提案手法の概要を示す。入力は、テレビ番組映像から取得した 1 枚のフレーム画像である。まず、走査窓の位置を少しずつ変化させながらフレーム全体を走査し、その窓領域が字幕テキストの一部であるかを判定することで候補領域を求める。判定には SVM を利用する。続いて、候補領域と判定された走査窓が、一定数以上重なっている領域を抽出する。最後に、各領域の外接矩形を求め、矩形内のエッジ密度や矩形サイズに基づく判定の後、検出結果を出力する。以降、各処理の詳細について説明する。

3.1 エッジ画像の生成

まず、提案手法で利用するエッジ特徴について説明する。テレビ番組では、色の差異によって文字の輪郭を強調する場合があるため、RGB 画像から特徴を抽出する。フレーム画像の座標 (x, y) 、チャネル c ($c \in \{R, G, B\}$) における x 方向および y 方向の勾配を $G_{c,x}(x, y)$, $G_{c,y}(x, y)$ としたとき、座標 (x, y) における x 方向および y 方向の勾配 $G_x(x, y)$, $G_y(x, y)$ を以下のように定める。

$$G_x(x, y) = \max_{c \in \{R, G, B\}} G_{c,x}(x, y) \quad (1)$$

$$G_y(x, y) = \max_{c \in \{R, G, B\}} G_{c,y}(x, y) \quad (2)$$

ここで $G_{c,x}(x, y)$, $G_{c,y}(x, y)$ は Sobel フィルタを用いて算出する。フレーム画像の座標 (x, y) におけるエッジ強度 $G(x, y)$ およびエッジ方向 $\theta(x, y)$ は次式で与えられる。

$$G(x, y) = \sqrt{G_x^2(x, y) + G_y^2(x, y)} \quad (3)$$

$$\theta(x, y) = \tan^{-1} \frac{G_y(x, y)}{G_x(x, y)} \quad (4)$$

3.2 走査窓による候補領域の検出

走査窓による候補領域の検出処理について述べる。図 2 に、走査窓の領域が字幕テキストかどうかを判定するための手順を示す。まず、計算コスト削減のため、走査窓内のエッジ密度に基づく判定を実施する。字幕テキスト領域には文字列を構成するエッジ成分が一定以上、存在している。そこで、走査窓内のエッジ密度が閾値 T_{sd} 以下である場合は、候補領域ではないと判定し、処理を終了する。走査窓領域 S のエッジ密度 $D(S)$ の算出式は以下のとおりである。

$$D(S) = \frac{1}{|S|} \sum_{(x, y) \in S} G(x, y) \quad (5)$$

図 2 走査窓による候補領域の検出手順 (図 1 (b) の処理)

図 3 学習に用いた画像データの例 (16 × 16 画素)

一方、エッジ密度 $D(S)$ が閾値 T_{sd} より大きい場合は、機械学習による判定を実施する。具体的には、走査窓内の画像領域から画像特徴ベクトルを算出し、SVM によって字幕テキスト領域の一部であるかを判定する。特徴ベクトルは、本論文で提案するエッジ空間パターンに加え、エッジ方向ヒストグラム²⁶⁾、エッジ強度ヒストグラム、ローカルバイナリパターン²⁷⁾、ハールウェーブレットテクスチャの合計 5 種類の画像特徴量を連結して求める。これらの画像特徴量は、従来研究において有用性が認められているものの中から選択した。いずれも走査窓内における文字の位置や数に依存しない特徴量となっている。また、SVM の学習には、文字の境界を無視して切り出した画像を利用する。図 3 に学習に用いた画像の例を示す。以下、各特徴量の詳細について述べる。

(1) エッジ空間パターン

エッジ空間パターンは、文字を形成する線や点の配置を反映した特徴量である。着目画素の周囲におけるエッジの分布パターンを解析することで、従来手法では検出が困難であったエッジの交差や直線以外のエッジにも対応できるようにする。また、エッジそのものではなく、エッジとエッジの間の空間の配置に着目することで、背景画像の変動への頑健性を確保する。具体的な算出手順を説明する。まず、

図 4 エッジ空間パターン $ESP_{8,1}$ の算出例 ($n=1$) 図 5 エッジ空間パターン $ESP_{8,2}$ のパターン一覧 ($n=2$)

式 (3) で求めたエッジ強度 G を閾値 T_{bi} で二値化し、各画素に「エッジ」または「非エッジ」のラベルを付与する。次に、着目画素 (x, y) と、その周囲 8 近傍の画素における任意の n 画素分のラベルに基づいて特徴ベクトル $\mathbf{ESP}_{8,n}$ を算出する。図 4 に $n=1$ の場合の特徴ベクトルの算出例を示す。着目画素がエッジ画素の場合、8 近傍における特定の 1 画素が非エッジ画素であるか否かに応じて、対応するベクトル要素に 1 または 0 を設定する。近傍 8 画素から 1 画素を選ぶ方法は全部で 8 通りであるため、8 次元のベクトルとなる。これに、着目画素が非エッジ画素である場合に 1 を設定する 1 次元を加えて、合計で 9 次元の特徴ベクトルを算出する。これを着目画素 (x, y) におけるエッジ空間パターン $\mathbf{ESP}_{8,1}(x, y)$ とする。図 4 の例では、着目画素の近傍におけるエッジおよび非エッジ画素の位置関係を 9 種類のパターンで表している。入力画像がパターンにマッチすれば、対応するベクトル要素を 1 に設定する。図 5 に、 $n=2$ の場合の各ベクトル要素に対応するパターンを示す。周囲の 8 画素から 2 画素を選ぶ方法は合計 28 通りであるため、 $\mathbf{ESP}_{8,2}(x, y)$ は 29 次元となる。同様に、任意の n ($n=1, \dots, 8$) に対応するエッジ空間パターン $\mathbf{ESP}_{8,n}$ が算出できる。

提案手法では、走査窓内の各画素について $n=1, \dots, 5$ の $\mathbf{ESP}_{8,n}$ を求め、それらの平均ベクトル $\overline{\mathbf{ESP}}_{8,1}, \dots, \overline{\mathbf{ESP}}_{8,5}$ を連結して特徴量とする。

(2) エッジ方向ヒストグラム

エッジ方向ヒストグラム²⁶⁾は、領域内におけるエッジ方向の関係に基づく特徴量である。式 (4) で求められるエッ

ジ方向 θ を k 段階に量子化し、各エッジ方向 k のエッジ強度 $\psi_k(x, y)$ を求める²⁸⁾。

$$\psi_k(x, y) = \begin{cases} G(x, y), & \text{if } \theta(x, y) \in \text{bin}_k^{(\theta)} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

走査窓領域 $\mathbf{S}$ におけるエッジ方向 k の累積エッジ強度 $E_k(\mathbf{S})$ は以下のように算出できる。

$$E_k(\mathbf{S}) = \sum_{(x,y) \in \mathbf{S}} \psi_k(x, y) \quad (7)$$

エッジ方向ヒストグラム $EOH_{k1,k2}(\mathbf{S})$ は、方向 k_1 と方向 k_2 の累積エッジ強度の比に基づいて定義する。

$$EOH_{k1,k2}(\mathbf{S}) = \frac{E_{k1}(\mathbf{S}) + \epsilon}{E_{k2}(\mathbf{S}) + \epsilon} \quad (8)$$

ここで ϵ は、0 による除算を避けるための係数である。

提案手法では、エッジ方向を 8 方向に量子化し、64 通りの $EOH_{k1,k2}$ ($k_1=1, \dots, 8$, $k_2=1, \dots, 8$) を算出する。

(3) エッジ強度ヒストグラム

エッジ強度ヒストグラムは、領域内におけるエッジ強度の分布に基づく特徴量である。式 (3) で求められるエッジ強度 G を k 段階に量子化し、走査窓領域 $\mathbf{S}$ における各エッジ強度 k の出現頻度 $EMH_k(\mathbf{S})$ を求める。

$$\varphi_k(x, y) = \begin{cases} 1, & \text{if } G(x, y) \in \text{bin}_k^{(G)} \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

$$EMH_k(\mathbf{S}) = \frac{1}{|\mathbf{S}|} \sum_{(x,y) \in \mathbf{S}} \varphi_k(x, y) \quad (10)$$

提案手法では、エッジ強度を 16 段階に量子化して特徴量 $EMH_k(\mathbf{S})$ ($k=1, \dots, 16$) を算出する。

(4) ローカルバイナリパターン

ローカルバイナリパターン²⁷⁾は、着目画素と周辺画素との輝度値の大小のパターンに基づく特徴量である。半径 r の円周上の p 個の画素から算出されるローカルバイナリパターン $LBP_{p,r}$ は以下のように定義される。

$$LBP_{p,r}(x, y) = \begin{cases} \sum_{p=0}^{P-1} \delta_{p,r}(x_i, y_i), & \text{if } U_{p,r}(x, y) \leq 2 \\ p+1, & \text{otherwise} \end{cases} \quad (11)$$

ここで、 $\delta_{p,r}$ は着目画素 (x, y) と周辺画素 $(x+x_i, y+y_i)$ との輝度値の大小関係を表し、次のように算出される。

$$\delta_{p,r}(x_i, y_i) = \begin{cases} 1, & f(x+x_i, y+y_i) - f(x, y) \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

式中の x_i, y_i ($0 \leq i \leq p-1$) は以下で与えられる。

$$x_i = r \cos \frac{2\pi i}{p}, \quad y_i = r \sin \frac{2\pi i}{p} \quad (13)$$

また、式 (11) における $U_{p,r}$ は、 $\delta_{p,r}$ の並びにおいて 0 と 1 が変化する箇所の総数を表す。

表 1 走査窓の重なりによる字幕候補領域の判定手順
アルゴリズム: 走査窓の重なりによる判定

```

F: フレーム画像全体の画素
for all  $(x, y) \in \mathbf{F}$  do  $s(x, y) \leftarrow 0$ 
while フレーム全体の走査が終わるまで do begin
  走査窓  $\mathbf{S}$  の位置をずらす
  if  $\mathbf{S}$  が字幕テキストの候補領域 then begin
    for all  $(x, y) \in \mathbf{S}$  do  $s(x, y) \leftarrow s(x, y) + 1$ 
  end
end

for all  $(x, y) \in \mathbf{F}$  do begin
  if  $s(x, y) < T_{ov}$  then 画素  $(x, y)$  を除外する
end.

```

$$U_{p,r}(x, y) = |\delta_{p,r}(x_{p-1}, y_{p-1}) - \delta_{p,r}(x_0, y_0)| + \sum_{i=1}^{p-1} |\delta_{p,r}(x_i, y_i) - \delta_{p,r}(x_{i-1}, y_{i-1})| \quad (14)$$

以上の処理によって $LBP_{p,r}$ を算出する。

提案手法では、走査窓内の各画素に対して $LBP_{p,r}$ を求め、その頻度ヒストグラムを特徴量とする。解像度変化に耐性を持たせるため、文献 27) と同様、 $(p, r) = (8, 1), (16, 2), (24, 3)$ のそれぞれについて $LBP_{p,r}$ の頻度ヒストグラムを算出する。

(5) ハールウェーブレットテクスチャ

ハールウェーブレットテクスチャは、画像の周波数特徴に基づく特徴量である。まず、走査窓内の画像領域に対してハールウェーブレット変換を 3 段階適用する。次に、それぞれのサブバンド領域における輝度値の分散を算出し、それらを連結して特徴量とする。

以上、(1) から (5) までの特徴量を連結することで、走査窓内の画像領域に対する特徴ベクトルとする。これを SVM で識別し、字幕テキストの候補領域を判定する。

3.3 走査窓の重なりによる判定

検出された走査窓の重なりに基づいて候補領域を絞り込む。字幕テキストは通常、ある程度の大きさをもってフレーム画像上に表示される。字幕テキスト領域のサイズに対して走査窓が十分に小さければ、字幕テキスト領域と判定される走査窓は図 1 (b) のように重なって検出されることになる。一方、誤って検出される走査窓については、単独あるいは少数で散在する傾向がある。これに基づいて誤検出の除外を試みる。具体的な手順を表 1 に示す。走査窓の重なり数 $s(x, y)$ を画素単位で保持しておき、フレーム全体の走査が終わった段階で $s(x, y)$ が閾値 T_{ov} より小さい画素を誤検出として除外する。

3.4 エッジ密度と矩形サイズによる判定

最後に、エッジ密度と矩形サイズに基づく判定を実施する。まず、3.3 で求めた画素群を領域単位に分割した後、各領域について外接矩形を検出する。次に、外接矩形内のエッジ密度と外接矩形のサイズを求め、それらが閾値より大きい領域を字幕テキストと判定する。以下に、外接矩形領域 $\mathbf{R}$ に対する判定条件を示す。

表 2 提案手法における閾値の設定

閾値	設定値	閾値	設定値
T_{sd}	380.0	T_{bi}	600.0
T_{ov}	2.0	T_{rd}	380.0
T_{ra}	700.0	T_{rw}	0.0
T_{rh}	0.0		

$$(D(\mathbf{R}) > T_{rd}) \text{ and } (W_{\mathbf{R}} \times H_{\mathbf{R}} > T_{ra}) \\ \text{and } (W_{\mathbf{R}} > T_{rw}) \text{ and } (H_{\mathbf{R}} > T_{rh}) \quad (15)$$

ここで、 $W_{\mathbf{R}}$ と $H_{\mathbf{R}}$ は、それぞれ外接矩形領域 $\mathbf{R}$ の幅と高さを表し、 $D(\mathbf{R})$ は外接矩形領域 $\mathbf{R}$ 内のエッジ密度を表す。 T_{rw} , T_{rh} , T_{ra} , T_{rd} は閾値である。エッジ密度 $D(\mathbf{R})$ の算出式は式 (5) と同様である。

以上の処理によって、入力フレーム画像から字幕テキスト領域の外接矩形を検出する。

4. 評価実験

提案手法の有効性を検証するため、実際に放送されたテレビ番組映像を利用した評価実験を実施した。映像のフレームサイズは横縦 432×240 画素である。フレーム画像は R, G, B の 3 チャンネルからなり、各画素はチャンネルごとに 0 から 255 までの値を持つものとする。提案手法の設定については、走査窓のサイズを 16×16 画素とし、水平垂直方向に 4 画素ずつ移動させながらフレーム全体を走査することとした。また、各種閾値については表 2 に示す設定を利用した。これらの設定は、処理時間と検出精度のバランスや、予備実験の結果などを考慮して決定した。SVM については、LIBLINEAR²⁹⁾を利用した。

4.1 走査窓による判別精度

始めに、本論文で提案したエッジ空間パターンの有効性を検証するため、走査窓による判別処理の精度を評価した。

(1) 実験の設定

実験には NHK で放送された 10 番組分の映像を利用した。映像の合計長は約 320 分である。番組のジャンルはニュースやドキュメンタリなど様々である。各番組映像から 1 秒ごとに 1 枚のフレーム画像を抽出し、各フレーム画像から走査窓と同様のサイズの部分画像をランダムに切り出した。切り出しの際には、3.2 の処理に合わせて、エッジ密度が閾値 T_{sd} より大きい領域のみを選択するようにした。切り出した画像の総数は、字幕テキスト領域（正例）が 9,500 枚であり、それ以外の領域（負例）が 9,500 枚である。そのうち、正例 7,000 枚と負例 7,000 枚を学習データとして利用し、残りの正例 2,500 枚と負例 2,500 枚をテストデータとして利用した。図 3 に実験に用いた学習データの例を示す。評価指標には以下に示す判別精度を用いた。

$$\text{判別精度} = \frac{\text{正しく判別されたテストデータの数}}{\text{テストデータの総数}} \quad (16)$$

(2) 実験結果

3.2 で提案したエッジ空間パターン (ESP), エッジ方向

表 3 特徴量の組合せによる判別精度の比較

#	特徴量					判別精度
	ESP (提案特徴)	EOH	EMH	LBP	HWT	
1	✓					0.813 (4066/5000)
2		✓				0.714 (3570/5000)
3			✓			0.700 (3500/5000)
4				✓		0.786 (3929/5000)
5					✓	0.798 (3988/5000)
6	✓	✓				0.778 (3890/5000)
7	✓		✓			0.849 (4246/5000)
8	✓			✓		0.874 (4369/5000)
9	✓				✓	0.884 (4422/5000)
10		✓	✓			0.753 (3763/5000)
11		✓		✓		0.861 (4307/5000)
12		✓			✓	0.795 (3975/5000)
13			✓	✓		0.824 (4121/5000)
14			✓		✓	0.875 (4373/5000)
15				✓	✓	0.795 (3976/5000)
16	✓	✓	✓			0.830 (4150/5000)
17	✓	✓		✓		0.917 (4585/5000)
18	✓	✓			✓	0.867 (4335/5000)
19	✓		✓	✓		0.866 (4330/5000)
20	✓		✓		✓	0.922 (4611/5000)
21	✓			✓	✓	0.889 (4447/5000)
22		✓	✓	✓		0.912 (4560/5000)
23		✓	✓		✓	0.902 (4510/5000)
24		✓		✓	✓	0.873 (4363/5000)
25			✓	✓	✓	0.837 (4187/5000)
26	✓	✓	✓	✓		0.930 (4650/5000)
27	✓	✓	✓		✓	0.904 (4522/5000)
28	✓	✓		✓	✓	0.925 (4625/5000)
29	✓		✓	✓	✓	0.893 (4467/5000)
30		✓	✓	✓	✓	0.912 (4562/5000)
31	✓	✓	✓	✓	✓	0.941 (4706/5000)

ヒストグラム (EOH), エッジ強度ヒストグラム (EMH), ローカルバイナリパターン (LBP), ハールウェーブレットテクスチャ (HWT) のすべての組合せのパターンについて, 判別精度を比較評価した。

表 3 に実験結果を示す。表 3 の 1 から 5 行目は各特徴量の個別の精度を表している。この中では, 本提案の ESP が 81.3% と最も高い精度となった。6 から 15 行目は 5 種類の特徴量のうち 2 種類を利用した場合である。この中で最も精度が高くなったのは本提案の ESP と, HWT の組合せであり, 判別精度は 88.4% となった。2 種類の特徴量を組合せることで, 個別の特徴量よりも判別精度が向上した。16 から 25 行目は 3 種類の特徴量を利用した場合である。この中で最も高い精度を達成したのは ESP, EMH, HWT の組であり, 判別精度は 92.2% であった。2 種類の場合よりもさらに精度が向上した。同様に 26 から 30 行目は 5 種類のうち 4 種類を利用した場合であり, HWT を除く 4 種類を使った場合が最も高い精度となった。判別精度は 93.0% であった。最後に 31 行目はすべての特徴量を使用した場合である。判別精度は 94.1% となり, 5 種類すべての特徴量を利用した場合が, すべての組合せパターンの中で最も高い精度となった。

次に, 従来の画像特徴量との比較実験を実施した。比較には, Wang らの画像特徴量²⁰⁾を用いた。Wang らの手法

表 4 従来手法との判別精度の比較

手法	判別精度
提案手法	0.941 (4706/5000)
Wang ら ²⁰⁾	0.735 (3673/5000)

図 6 Wang ら²⁰⁾の画像特徴量算出における領域分割

は, 文字領域を図 6 のような 8 領域に分割し, 各領域から輝度のコントラスト²⁰⁾と, 8 方向エッジ方向ヒストグラムを算出するものである。8 方向エッジ方向ヒストグラムは, 各画素のエッジ方向を 8 方向に量子化し, その頻度ヒストグラムを求める方式であり, 提案手法の EOH とは算出方法が異なる。比較の結果を表 4 に示す。提案手法の判別精度は 94.1% である。それに対して, Wang らの手法は 73.5% となり, 判別精度が低下する結果となった。輝度コントラストと 8 方向エッジ方向ヒストグラムの組合せでは, 字幕テキストと複雑な背景画像を精度よく判別することが難しいものと考ええる。

以上の結果から, 走査窓による判別においては, ESP, EOH, EMH, LBP, HWT の 5 種類の組合せが有効であることが確認できた。

4.2 字幕テキストの検出精度

続いて, 実際の放送番組映像を対象として, 字幕テキスト領域の検出精度を評価した。

(1) 実験の設定

実験には NHK で放送された 23 番組分の映像を利用した。番組のジャンルは様々であり, 映像の合計長は約 10 時間である。これらは, 4.1 の実験で利用した番組とは別のデータとなっている。それぞれの番組映像から 1 秒ごとに 1 枚のフレーム画像を抽出し, 提案手法への入力とした。抽出したフレーム画像の総数は 37,457 フレームである。それらに人手で字幕テキスト領域の外接矩形を付与して正解データとした。正解データとして得られた字幕テキスト領域の総数は 23,216 であった。SVM の学習には, 前節の実験と同様の学習データを利用した。検出精度の評価には, 以下に示す再現率と適合率, F 値を利用した。

$$\text{再現率} = \frac{N_{crr}}{N_{crr} + N_{miss}} \quad (17)$$

$$\text{適合率} = \frac{N_{crr}}{N_{crr} + N_{err}} \quad (18)$$

$$F \text{ 値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \quad (19)$$

ここで, N_{crr} は提案手法で正しく検出された字幕テキスト領域の数, N_{miss} は提案手法で未検出となった字幕テキスト領域の数, N_{err} は提案手法で誤検出された字幕テキスト領域の数を表す。つまり, 再現率は未検出の少なさを表す指標であり, 適合率は誤検出の少なさを表す指標といえる。F 値は, トレードオフの関係にある再現率と適合率を統合

図 7 画素単位から矩形単位の検出結果への変換方法

表 5 字幕テキスト領域の検出精度

番組 #	ジャンル	番組長	字幕 テキスト数	提案手法			Epshtein ¹⁹⁾			Washio ⁸⁾			Minetto ²¹⁾		
				再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
1	ドキュメンタリ	5 分	125	0.992	0.932	0.961	0.128	0.232	0.165	0.624	0.538	0.578	0.304	0.905	0.455
2	ニュース	45 分	3720	0.967	0.962	0.964	0.577	0.725	0.643	0.783	0.930	0.850	0.824	0.957	0.885
3	ドラマ	15 分	80	0.913	0.646	0.756	0.525	0.111	0.183	0.488	0.133	0.209	0.313	0.694	0.431
4	スポーツ	5 分	76	0.776	1.000	0.874	0.539	0.788	0.641	0.539	0.281	0.369	0.500	0.927	0.650
5	情報	25 分	371	0.933	0.576	0.712	0.294	0.061	0.101	0.682	0.583	0.629	0.709	0.780	0.743
6	趣味・教育	25 分	467	0.951	0.867	0.907	0.210	0.217	0.214	0.657	0.592	0.623	0.291	0.850	0.434
7	情報	25 分	754	0.520	0.812	0.634	0.285	0.239	0.260	0.560	0.420	0.480	0.418	0.741	0.534
8	ドキュメンタリ	5 分	136	0.897	0.819	0.856	0.125	0.191	0.151	0.632	0.589	0.610	0.375	0.927	0.534
9	ニュース	20 分	2190	0.928	0.978	0.952	0.248	0.524	0.337	0.862	0.808	0.834	0.752	0.976	0.849
10	スポーツ	5 分	58	0.569	1.000	0.725	0.431	0.397	0.413	0.448	0.100	0.163	0.638	1.000	0.779
11	スポーツ	120 分	4525	0.845	0.927	0.884	0.102	0.266	0.148	0.805	0.664	0.728	0.428	0.944	0.589
12	ニュース	30 分	2398	0.984	0.966	0.975	0.123	0.307	0.176	0.847	0.699	0.766	0.748	0.961	0.841
13	ニュース	60 分	4394	0.952	0.964	0.958	0.312	0.525	0.392	0.743	0.861	0.798	0.737	0.977	0.840
14	ドキュメンタリ	45 分	1243	0.901	0.611	0.728	0.240	0.172	0.201	0.697	0.539	0.608	0.443	0.764	0.561
15	趣味・教育	15 分	266	0.703	0.770	0.735	0.162	0.098	0.122	0.421	0.477	0.447	0.711	0.836	0.768
16	バラエティ	40 分	867	0.847	0.679	0.754	0.054	0.038	0.045	0.683	0.277	0.394	0.670	0.949	0.786
17	ドキュメンタリ	15 分	274	0.982	0.815	0.891	0.139	0.091	0.110	0.620	0.480	0.541	0.803	0.991	0.887
18	スポーツ	10 分	92	0.967	0.669	0.791	0.467	0.589	0.521	0.511	0.228	0.315	0.478	0.299	0.368
19	福祉	25 分	241	0.797	0.985	0.881	0.195	0.119	0.148	0.743	0.108	0.188	0.398	0.681	0.503
20	ドラマ	30 分	11	0.727	0.042	0.080	0.182	0.002	0.004	0.364	0.005	0.009	0.364	0.078	0.129
21	福祉	5 分	42	0.976	1.000	0.988	0.333	0.250	0.286	0.690	0.145	0.240	0.786	1.000	0.880
22	趣味・教育	30 分	546	0.577	0.726	0.643	0.405	0.159	0.228	0.566	0.407	0.473	0.581	0.748	0.654
23	ドキュメンタリ	25 分	340	0.932	0.734	0.821	0.559	0.275	0.369	0.618	0.543	0.578	0.565	0.865	0.683
	合計	625 分	23216	0.899	0.880	0.889	0.273	0.308	0.289	0.754	0.603	0.670	0.638	0.923	0.754

した指標である．いずれの指標も数値が大きいほど精度が高いことを表す．

(2) 比較手法

提案手法の有効性を検証するため、従来手法との比較を実施した．比較手法としては、伝統的な字幕テキスト検出手法である鷲尾らの手法⁸⁾ (Washio と記述) と、最新の手法の中から Epshtein らの手法¹⁹⁾ (Epshtein と記述) と Minetto らの手法²¹⁾ (Minetto と記述) を選択した．Epshtein は、SWT と呼ばれるシーンテキスト検出の代表的な手法であり、文字のストローク幅に基づいてテキスト領域を判定する手法である．本実験では Kumar らによる実装³⁰⁾ を利用した．Washio は、静止したエッジの密集に基づいて字幕テキスト領域を判定する手法である．Minetto は、SnooperText と呼ばれるシーンテキストの検出手法であり、フーリエ変換などに基づく特徴量を SVM で判別してテキスト領域を検出する手法である．本実験では、Minetto による実装³¹⁾ を利用した．

比較手法のうち Epshtein と Washio は、テキスト領域を画素単位で検出する方式となっている．本実験では、矩形単位で検出精度を評価するため、これらの検出結果を矩形単位に変換した上で評価する．変換手順を図 7 に示す．ま

ず、検出された画素に走査窓を割り当てる．その後、提案手法と同様に、走査窓の重なりに基づく判定と、エッジ密度と矩形サイズに基づく判定を実施し、矩形単位の検出結果に変換する．

また精度評価において、字幕テキスト以外は誤検出とするのが本来であるが、比較手法のうち Epshtein と Minetto はシーンテキストも検出できる手法となっている．実験データには大量のシーンテキストが含まれており、事前にすべてを取り除いておくことは困難なため、本実験ではシーンテキストを誤検出とカウントせずに検出精度を算出することとした．なお、シーンテキストを対象としていない提案手法と Washio については、シーンテキストを誤検出とカウントした場合についても別途、検出精度を評価する．

(3) 実験結果

実験結果を表 5 に示す．評価の結果、提案手法の合計の精度は再現率が 89.9%、適合率が 88.0%、F 値が 0.889 となった．再現率および F 値については、比較手法の中で最も高い結果となった．番組ごとの比較でも、提案手法の F 値が 23 番組のうち 17 番組において最も高い結果となった．ほとんどすべての番組で F 値が 0.7 以上という良好な結果となり、ニュースやドキュメンタリ、バラエティなどの

図 8 提案手法による正検出の例

図 9 提案手法による誤検出の例

図 10 提案手法による未検出の例

様々な映像に対して頑健に動作することが確認できた。特にニュース番組においては、F 値が 0.95 以上という非常に高い精度を達成することができた。ニュース番組の字幕テキストは、背景領域との境界が明瞭なものが多かったため、検出精度が向上したものと考える。一方、字幕テキストの表示が少なかった#20 のドラマ番組については、11 個の正解に対して 182 個の誤検出が発生し、適合率が大きく低下する結果となった。誤検出は、出演者の服に描かれた細かな絵柄や、人ごみのシーンなどで多く起きていた。

提案手法に次いで F 値が高くなったのは Minetto であった。再現率は 63.8%、適合率は 92.3%、F 値は 0.754 であった。適合率は比較手法の中でも最も高い結果となった。Minetto では、通常の判定処理に加えて、誤検出軽減のための後処理を実施しており、これによって適合率が向上したものと考えられる。しかし、厳しい判定によって未検出が増加し、全般的に再現率が低下する結果となった。そのため、F 値による比較では、提案手法よりも 0.135 低い結果となった。

Washio については、再現率が 75.4%、適合率が 60.3%、F 値が 0.670 という結果になった。エッジの密集という簡易な特徴に基づいているため、周波数の高い領域などで誤検出が多く発生し、適合率が低下する結果となった。また、同じ位置に出現し続けるエッジを字幕テキストの候補としているため、表示位置や文字サイズが時間的に変化するようなものは検出できなかった。

Epshtein については、再現率が 27.3%、適合率が 30.8%、F 値が 0.289 となった。複雑な背景に重畳される字幕テキストから文字のストローク幅を正確に検出することは難しく、精度が大きく低下する結果となった。また、Kumar による実装³⁰⁾は、携帯端末で撮影された画像に適合するように改変されており、この点も精度低下の一因と考えられる。具体的な改変内容は、エッジペアの検出、および連結成分の抽出に関わる閾値が携帯端末の画質や画像サイズに合わせて調整されている点と、連結成分のフィルタリングに関わる処理の一部が省略されている点である。

提案手法と Washio について、シーンテキストの検出を誤検出とカウントした場合の評価結果を示す。提案手法については、再現率が 89.9%、適合率が 83.9%、F 値が 0.868 となり、適合率が約 4% 低下した。それでもなお、比較手法の中で F 値が最も高い結果となっている。Washio については再現率が 75.4%、適合率が 52.2%、F 値が 0.617 となり、適合率が約 8% 低下した。

提案手法の検出結果の詳細を述べる。図 8 に提案手法による正検出の例を示す。文字の色やサイズ、フォントの種類、テキストの方向、出現位置、背景などによらず、様々なタイプの字幕テキストが正確に検出できていることがわかる。本論文で提案した画像特徴量と走査窓による判定処理によって、字幕テキストとそれ以外の領域とを精度よく判別できたものとする。図 9 に提案手法による誤検出の例を示す。格子状の領域や、背景との明度差が大きい高周波領

表 6 各処理段階 (図 1) における字幕テキスト領域の検出精度

処理段階	検出精度		
	再現率	適合率	F 値
(b) 走査窓による候補領域の検出	0.971	0.376	0.543
(c) 走査窓の重なりによる判定	0.968	0.435	0.601
(d) エッジ密度と矩形サイズによる判定	0.899	0.880	0.889

表 7 処理時間の比較

手法	処理時間	
	1 フレームあたり	1 時間 (3600 フレーム) あたり
提案手法	0.275 秒	約 16 分 30 秒
Epshtein ¹⁹⁾	0.133 秒	約 7 分 58 秒
Washio ⁸⁾	0.050 秒	約 3 分 00 秒
Minetto ²¹⁾	21.239 秒	約 21 時間 14 分 20 秒

域において誤検出が発生した。また、イラストなどにおける線の集合を文字と誤検出する場合があった。判定処理に用いる特徴量を増やしたり、SVM の学習データの内容や、学習アルゴリズムの調整が必要と考える。また、Minetto のような誤検出の軽減のための後処理を追加するといった検討が必要と考える。図 10 に提案手法による未検出の例を示す。文字サイズが極端に大きいものや、文字が大きく変形しているものが未検出となった。また、半透明の字幕テキストが未検出となる場合があった。文字サイズについては、走査窓による判定の際に、複数のスケールで走査することで対応が可能と考える。半透明の字幕テキストについては、エッジ特徴に基づく判定が困難であるため、別のアプローチを検討する必要がある。例えば、原画像から直接、文字らしさを判定するような方法が考えられる。

(4) 各処理段階における検出精度

提案手法における各処理の寄与が明確となるよう、各段階における検出精度を評価した。実験結果を表 6 に示す。表 6 の各行は、図 1 に示す提案手法の各処理が終了した段階での字幕テキストの検出精度を表している。実験の結果、図 1 (b), (c), (d) の各処理が完了した段階での F 値は、それぞれ 0.543, 0.601, 0.889 となった。処理の進行にしたがって、適合率と F 値が段階的に向上している。特に、図 1 (d) の判定処理による適合率の向上が顕著となった。一方、再現率の低下は 1% 以内となっており、各段階において誤検出が効果的に取り除かれていることが確認できた。

(5) 処理時間の比較

表 7 に処理時間の比較結果を示す。実行環境は、CPU が Intel Xeon X5690 3.47 GHz で、メモリは 72 GB である。表 7 には、1 フレームあたりの平均処理時間と、1 時間分のフレーム画像の処理時間を示している。本実験では 1 秒につき 1 枚のフレーム画像を取得しているため、1 時間分は 3,600 フレームとなる。

実験の結果、提案手法の処理時間は 1 フレームあたり 0.275 秒、1 時間あたり約 16 分 30 秒となった。Epshtein と Washio については、提案手法よりも処理時間が短い結果となった。1 フレームあたりの処理時間は、Epshtein が 0.133 秒、Washio が 0.05 秒であり、それぞれ提案手法の

約 1/2, 約 1/6 の処理時間となった。これらの手法は、走査窓による判定処理のような計算コストの高い処理を内包しないため、提案手法よりも高速な処理が可能になったものとする。最後に Minetto については、1 フレームあたりの処理時間が 21.239 秒、1 時間あたりの処理時間が 21 時間以上という結果になった。提案手法の約 77 倍の処理時間であり、他の従来手法と比較しても特に長い結果となった。Minetto の処理時間が増大した原因としては、機械学習による判定処理を多段に実行している点が挙げられる。これは、誤検出軽減のための措置であるが、多数の候補領域に対して判定処理を実行するため計算コストの増加につながった。また、文字候補領域を検出するための画素の連結判定や、文字候補領域の位置関係から文字列候補を抽出する処理にも多くの計算時間が費やされている。大量の映像データを処理する場合は、検出精度だけでなく、計算コストも重要な要素の一つになると考える。提案手法においても今後、検出精度と計算コストのバランスを考慮しながら手法の改良をすすめたい。

5. む す び

本稿では、エッジ空間パターンを利用したテレビ番組映像からの字幕テキスト領域の検出手法を提案した。エッジ空間パターンは、文字を形成する線や点の配置を反映した特徴量であり、着目画素の周囲におけるエッジの分布に基づいて算出される。提案手法では、まず、フレーム画像を走査窓で走査し、窓領域からエッジ空間パターンを含む複数の特徴量を算出した。これをサポートベクターマシンで判別することによって字幕テキストの候補領域を検出した。その後、得られた候補領域をエッジ密度や領域のサイズなどに基づいて選別することで、字幕テキストの矩形領域を検出した。約 10 時間分の放送映像を対象とした実験では、再現率が 89.9%、適合率が 88.0%、F 値が 0.889 という良好な結果が得られ、従来手法よりも F 値が 0.135 向上することが確認できた。今後は、より高精度にテキスト領域を判別するための特徴量の追加や、誤検出軽減のための方策などについて検討をすすめたい。

〔文 献〕

- 1) S. Kiranyaz and M. Gabbouj: "Hierarchical Cellular Tree: An Efficient Indexing Scheme for Content-Based Retrieval on Multimedia Databases", IEEE Trans. Multimedia, **9**, 1, pp.102-119 (2007)
- 2) M.A. Smith and T. Kanade: "Video Skimming for Quick Browsing Based on Audio and Image Characterization", Tech. Report CMU-CS-95-186, Carnegie Mellon University (1995)
- 3) P. Viola and M.J. Jones: "Robust Real-Time Face Detection", Int. Journal of Computer Vision, **57**, 2, pp.137-154 (2004)
- 4) N. Dalal and B. Triggs: "Histograms of Oriented Gradients for Human Detection", in Proc. IEEE CVPR, **1**, pp. 886-893 (2005)
- 5) T. Sato, T. Kanade, E. K. Kughes, M. A. Smith and S. Satoh: "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions", ACM Multimedia Syst., **7**, 5, pp.385-395 (1999)
- 6) N. Babaguchi, Y. Kawai and T. Kitahashi: "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration",

- IEEE Trans. Multimedia, **4**, 1, pp.68–75 (2002)
- 7) K. Junga, K.I. Kim and A.K. Jain: “Text Information Extraction in Images and Video: A Survey”, Pattern Recognition, pp.977–997 (2004)
- 8) 鷲尾 誠一, 有木 康雄, 緒方 淳: “クロスメディア・パッセージ検索 – テロップや CG フリップ文字列を検索質問とした発話文書に対する検索方式 –, 信学論, **J84-D-II**, 8, pp.1809–1816 (2001)
- 9) 佐藤 隆, 新倉 康巨, 谷口 行信, 阿久津 明人, 外村 佳伸, 浜田 洋: “MPEG 符号化映像からの高速テロップ領域検出法”, 信学論, **J81-D-II**, 8, pp.1847–1855 (1998)
- 10) 藤本 雅清, 有木 康雄, 松本 宏: “音声情報と画像情報の併用による商品紹介映像のセグメンテーション”, 信学論, **J89-D**, 2, pp.292–304 (2006)
- 11) 新井 啓之, 桑野 秀豪, 倉掛 正治, 杉村 利明: “映像中のテロップ表示フレーム検出方法”, 信学論, **J83-D-II**, 6, pp.1477–1486 (2000)
- 12) C. Garcia and X. Apostolidis: “Text Detection and Segmentation in Complex Color Images”, in Proc. IEEE Int. Conf. on ASSP, pp.2326–2329 (2000)
- 13) V. Wu, R. Manmatha and E.M. Riseman: “Finding Text in Images”, in Proc. ACM Int. Conf. on Digital Libraries, pp.23–26 (1997)
- 14) Y. Zhong, H. Zhang and A.K. Jain: “Automatic Caption Localization in Compressed Video”, IEEE Trans. Pattern Analysis and Machine Intelligence, **22**, 4, pp.385–392 (2000)
- 15) 河合 吉彦, 住吉 英樹, 藤井 真人: “単一映像フレームからの字幕領域検出手法”, 映メ冬大, 3-7 (2010)
- 16) R. Lienhart and A. Wernicke: “Localizing and Segmenting Text in Images and Videos”, IEEE Trans. Circuits and Syst. for Video Tech., **12**, 4, pp.256–268 (2002)
- 17) J. Zhang and R. Kasturi: “Text Detection Using Edge Gradient and Graph Spectrum”, in Proc. ICPR, pp.3979–3982 (2010)
- 18) H. Bai, J. Sun, S. Naoi, Y. Katsuyama, Y. Hotta and K. Fujimoto: “Video Caption Duration Extraction”, in Proc. ICPR, pp.1–4 (2008)
- 19) B. Epshtein, E. Ofek and Y. Wexler: “Detecting Text in Natural Scenes with Stroke Width Transform”, in Proc. IEEE CVPR, pp.2963–2970 (2010)
- 20) X. Wang, L. Huang and C. Liu: “A New Block Partitioned Text Feature for Text Verification”, in Proc. ICDAR, pp.366–370 (2009)
- 21) R. Minetto, N. Thome, M. Cord, N.J. Leite and J. Stolfi: “SnooperText: A Text Detection System for Automatic Indexing of Urban Scenes”, Computer Vision and Image Understanding, **122**, pp.92–104 (2014)
- 22) R. Minetto, N. Thome, M. Cord, N.J. Leite and J. Stolfi: “T-HOG: An Effective Gradient-based Descriptor for Single Line Text Regions”, Pattern Recognition, **46**, 3, pp.1078–1090 (2013)
- 23) X. Tang, X. Gao, J. Liu and H. Zhang: “A Spatial-Temporal Approach for Video Caption Detection and Recognition”, IEEE Trans. Neural Networks, **13**, 4, pp.961–971 (2002)
- 24) R.E. Schapire and Y. Singer: “Improved Boosting Algorithms Using Confidencerated Predictions”, Machine Learning, pp.297–336 (1999)
- 25) L. Neumann and J. Matas: “Real-Time Scene Text Localization and Recognition”, in Proc. IEEE CVPR, pp.3538–3545 (2012)
- 26) K. Levi and Y. Weiss: “Learning Object Detection from a Small Number of Examples: The Importance of Good Features”, in Proc. IEEE CVPR, **2**, pp.53–60 (2004)
- 27) T. Ojala, M. Pietikainen and T. Maenpaa: “Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns”, IEEE Trans. Pattern Analysis and Machine Intelligence, **24**, 7, pp.971–987 (2002)
- 28) 八木 康史, 斎藤 英雄 編, 藤吉 弘亘, 山下 隆義, 岡田 和典, 前田 英作, ノジク・ヴァンソン, 石川 尋代, ドゥソルビエ・フランソワ: “コンピュータビジョン 最先端ガイド 2”, アドコム・メディア株式会社 (2010)
- 29) R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin: “LIBLINEAR: A Library for Large Linear Classification”, Journal of Machine Learning Research, **9**, pp.1871–1874 (2008)
- 30) S. Kumar and A. Perrault: “Text Detection on Nokia N900 Using Stroke Width Transform”, <https://github.com/aperrault/DetectText> (2010)
- 31) R. Minetto: “SnooperText Source Code (Java)”, <http://www.dainf.ct.utfpr.edu.br/~rminetto/projects/snoopertext/>

かわい よしひこ
河合 吉彦 1999 年, 大阪大学基礎工学部情報工学科卒業。2001 年, 同大学大学院修士課程修了。同年, NHK 入局。放送技術局を経て, 2005 年から放送技術研究所勤務。博士 (工学)。メディア処理の研究に従事。正会員。

もちづき たかひろ
望月 貴裕 1994 年, 東京工業大学情報工学科卒業。1996 年, 同大学大学院修士過程修了。同年, NHK 入局。放送技術局を経て, 1998 年から放送技術研究所勤務。博士 (工学)。映像検索技術の研究開発に従事。正会員。

すみよし ひでき
住吉 英樹 1980 年, 広島県立広島工業高校電気科卒業。同年, NHK 入局。広島放送局を経て, 1984 年から放送技術研究所勤務。現在, 放送技術研究所 上級研究員。博士 (工学)。コンピュータを応用した番組制作システム, メタデータ制作システムの研究に従事。正会員。