

複数人物の表情認識による観客の満足度の自動定量評価

松 川 徹^{†1} 日高 章理 ^{†1} 栗田 多喜夫^{†2,†1}

スポーツやお笑い等のエンターテインメント産業では、観客が満足しているかいないかを評価することは、サービスの質を評価する上で重要である。通常では、観客の満足度を評価するためにアンケート調査が用いられているが、アンケート調査を行うためには、コストと手間が必要となる。そのため観客の満足度調査をビデオ映像から自動的に行うことが望まれている。本論文では、映像中の顔に着目して、観客の満足度を推定するシステムを提案する。提案システムは、顔識別器ベースの bag-of-visual-words に基づいており、顔表情の識別器を画像上に検出された全ての顔に適用することにより複数の人物の表情を認識し、表情と顔の向きを投票したヒストグラムを作る。このヒストグラムを特徴ベクトルとした識別器を用いて観客の満足度を自動的に推定する。サポートベクターマシンによるシーン判別を行い“喜んでいる”か“喜んでいない”、“真剣に観戦している”か“真剣に観戦していない”かの判別をある程度行うことができた。

Automatically Evaluation of Degree of Spectators' Satisfaction in Video Sequences by Recognizing Their Facial Expressions

TETSU MATSUKAWA,^{†1} AKINORI HIDAKA^{†1}
and TAKIO KURITA^{†2}

In this paper, we proposed a classification method of spectators' state in video sequences by voting of facial expressions and face directions. The task of this paper is to classify the state of the spectators in a given video sequence into “Positive Scene” or “Negative Scene”, and “Watching Seriously” or “Watching Not Seriously”. The proposed classifier is designed by a “bag-of-visual-words” approach based on face recognitions. First, the multiview (left-profile, front, right-profile) faces are detected from each image in the given video sequence. Then the detected faces are classified into the two expressions, smile or not smile. The classification results of the face directions and the facial expressions are voted to each classes' histogram over the video sequence. Finally, the state

of the spectators is classified by using the kernel SVM on the voted histograms. We conducted experiments using spectators' video sequences captured from TV. Our approach demonstrated promising results for classifications of “Positive Scene” and “Negative Scene” or “Watching Seriously” and “Watching Not Seriously”. It was also ascertained that the facial expression is important in the classification of “Positive” and “Negative”. On the other hand, face direction is important to classify whether the spectators are “Watching Seriously” or “Not”.

1. はじめに

スポーツやお笑い等のエンターテインメント産業では、観客が満足しているかいないかを評価することは、サービスの質を評価する上で重要である。通常では、観客の満足度を評価するためにアンケート調査が用いられているが、アンケート調査を行うためには、コストと手間が必要となる。そのため観客の満足度調査をビデオ映像から自動的に行うことが望まれている。そこで本論文では、映像中から観客の顔の有無と表情をパターン認識手法によって自動的に認識し、その認識結果に基づいて観客の満足度を推定するシステムを提案する。

現在、画像から顔を検出する手法は数多く提案されている。Viola らは、ブースティングにより特徴選択した弱識別器をカスケード型に接続し、インテグラルイメージを用いて Harr-like 特徴を高速に計算することにより高速かつ高精度に顔を検出する手法を提案した¹⁾。また、Haung らは、Viola らの手法を Width-First-Search tree structure と sparse features in granular space を用いて異なる向きの顔を検出する手法を提案した²⁾。また、Hidaka らは、Particle Swarm Optimization を用いることで、Viola らの手法の学習時間を高速化する手法を提案⁵⁾しているなど、Viola らの顔検出手法には数多くの拡張が提案されており、現在では顔を検出することは容易に行える状況である。

画像から顔表情を認識する研究も数多く行われている。Shinohara らは、フィッシャー判別基準に基づく局所特徴の最適重みマップを提案した⁶⁾。Hu らは、異なる向きの顔表情を認識するための数種類の局所特徴と次元圧縮法を比較した⁷⁾。Chen らは、選択された

^{†1} 筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻
Department of Computer Science, Graduate School of Systems and Information Engineering,
University of Tsukuba

^{†2} 産業技術総合研究所脳神経情報研究部門
Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology (AIST)

Harr-like 特徴と Gabor 特徴を用いて顔の検出と顔表情認識を一体化した⁸⁾。

このように、顔検出や表情認識の手法はこれまで数多く提案されているが、スポーツ観戦で多数の観客が試合を楽しんでいるかどうかなどといった群衆全体の状況を顔の情報に基づいて認識する研究はあまり行われていない。

そこで本研究では観客の顔の向きと表情に着目した満足度推定法を提案する。提案システムではまず顔検出器によって動画中から観客の顔を複数見つけだし、次に表情識別器によってそれらの顔の表情を認識する。この作業を顔の向き（左向き、正面向き、右向き）ごとに個別に行った後、認識された顔の向きと表情の頻度を数えあげてヒストグラム特徴を生成する。これは顔識別器（顔検出器 + 表情識別器）の識別結果に基づいて画像パッチを量子化することに相当し、すなわち提案特徴は顔識別器ベースの bag-of-visual-words 法に基づいているといえる。この提案特徴を元にサポートベクターマシン等の識別器を訓練することにより、シーンに含まれる観客の顔の向きと表情の頻度に応じて観客の満足度を自動的に推定するシステムを得ることができる。

近年、映像コンテンツを視聴するユーザの表情を解析することで、ユーザの映像コンテンツに対する関心度を推定する研究が行われている¹⁷⁾¹⁸⁾が、彼らのシステムは一人の人間が正面を見ている状況を仮定し、一人の人間の関心度を推定するものであるが、提案システムは、撮影状況を限定せず、多数の人間の表情から全体の満足度を推定する点異なる。

以下、2 章で提案システムの概要について説明し、3,4,5 章で提案手法の詳細について述べ、6 章でシステムの画面例を示す。7 章では実験結果を示し、8 章ではまとめを述べる。

2. 提案システムの概要

2.1 提案システムの扱う問題

提案システムは、観客映像に対して機械学習により満足度を推定するシステムである。提案システムを訓練するにはあらかじめ正解となる満足度が付与された学習用の映像データセット $\{x_i, y_i\}_{1:N}$ が必要となる (N はサンプル数)。ここで x_i は観客動画中の連続する M フレーム (1 フレームは $W \times H$ ピクセル) からなる画像シーケンスである (本研究では具体的には 30 フレーム、 720×480 ピクセルとしている)。また y_i は画像シーケンス x_i 内の観客が観戦対象に感じている“満足度”であり、訓練データの作成者が手動で与えるものである。

ここで画像シーケンスの満足度をどのように与えるかが問題となる。仮にその道の専門家などが訓練用シーケンス x_i を観て観客の満足度 y_i を 0.8, 0.2 などのように実数値で与



図 1 観客の画像例. (a): “喜んでいる”, (b): “喜んでいない” - “真剣に観戦していない”, (c): “喜んでいない” - “真剣に観戦している”. これらは説明用であり、実験には 2008 年北京オリンピックの観客映像を用いた。

Fig. 1 Example of spectators' scene. (a): “Positive Scene”, (b): “Negative Scene” - “Watching Not Seriously”, (c): “Negative Scene” - “Watching Seriously”. These are for explanation and not the actual Olympic scenes used in this paper.

えることが可能な場合、この問題を回帰問題の枠組みで取り扱うことが可能になる。またそのような複雑な定量化が難しいには、訓練用シーケンス x_i にして単純に“満足している ($y_i = +1$)”か“満足していない ($y_i = -1$)”かというラベルを割り当てる方法もある。この場合は満足度の推定問題を 2 クラスの識別問題として取り扱い、観客の満足度の度合いとしては識別時のスコア (識別境界からの距離など) を割り当てればよいと考えられる。

何をもって観客が満足していると見なすかは、ユーザーの目的や応用例によって異なるものと考えられる。本研究では簡単な例として、“観客が喜んでいる” = “観客が満足している”と見なす場合、および“観客が満足している” = “観客が真剣に観戦している”と見なす場合の 2 例を取り扱う。すなわち前者のケースでは観客が喜んでいる (と訓練データ作成者が判断したシーケンス) に満足度 $y_i = +1$ が割り当てられ、逆に観客が喜んでいないシーケンスには満足度 $y_i = -1$ が割り当てられる。同様に後者のケースでは観客が真剣に観戦しているシーケンスとそうでないシーケンスにそれぞれ満足度 $y_i = +1$ と $y_i = -1$ を割り当てる。図 1 に対象シーンの例を示す。満足度推定方法の詳細は 5.3 節に述べる。

本論文で実際に扱う観客動画は、スポーツを観戦している観客の動画である。観客動画は、2008 年北京オリンピックの TV 放映映像から、観客が映る場面を手動でサーチして切り出した映像を用いた。映像は 30fps のビデオレートで撮影されている。ここでは“観客が喜んでいる”に対応するデータとして、主に映っている観客の選手が勝ったときや優勢な状況の観客映像を 50 サンプル用意した。“喜んでいない”に対応するデータとして、競技が始まっていない時や主に映っている観客の国の選手が優勢でない状況の観客映像を 50 サンプル用意した。またこの“喜んでいない”データをさらに“真剣に観戦している”、“真剣に観戦していない”に分類する。“真剣に観戦している”に対応するデータは、主に映ってい

3 複数人物の表情認識による観客の満足度の自動定量評価

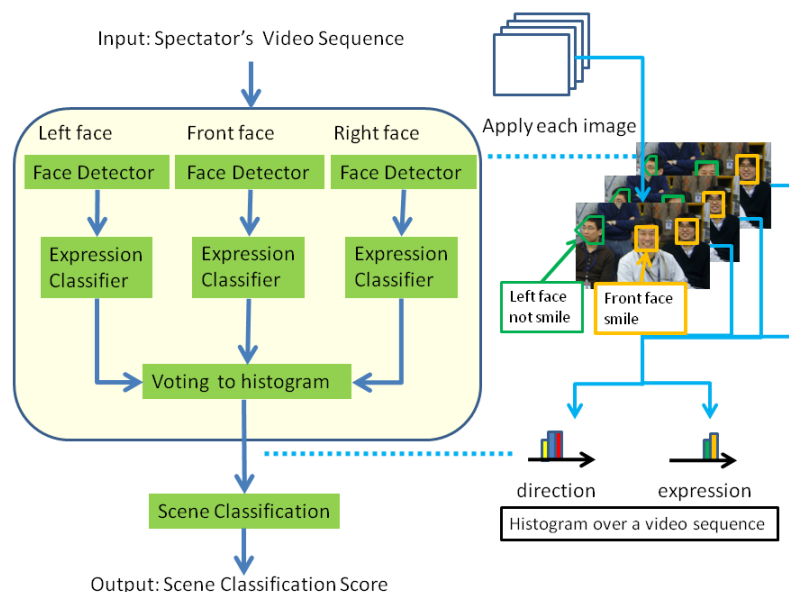


図 2 提案システムの構成概要
Fig. 2 Overview of proposed system.

る観客の国の選手が優勢でない状況の観客映像が 25 サンプルである。“真剣に観戦していない”に対応するデータは、競技が始まっていない時の観客映像が 25 サンプルである。顔検出・表情認識するときには、画像をグレースケールに変換する。それぞれのフレームには 3～30 人の観客が映っている。

2.2 提案システムの構成概要

提案システムでは、観客が“喜んでいる”か“喜んでいない”かを識別することに観客が笑っているかいないかが重要と考え、笑顔が非笑顔かの情報を利用する。また、“真剣に観戦している”か“真剣に観戦していない”かを識別するために観客の視線の集中度が重要と考え、映像中の顔の大まかな向きの頻度も利用する。提案システムの構成概要を図 2 に示す。初めに Viola らによって提案された顔検出手法によって、顔の向き毎（左向き、正面、右向き）の顔を観客動画の各フレームで検出する。次に検出された顔を笑顔、非笑顔の 2 種類の表情に識別する。1 つの画像シーケンスにおいて顔の向きと表情をそれぞれ

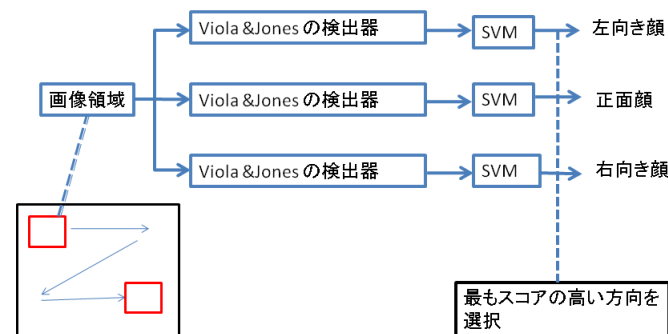


図 3 顔検出手法
Fig. 3 Face Detection Method.

のヒストグラムに投票する。投票されたヒストグラムを特徴ベクトルに用いたサポートベクターマシン (SVM) により、各状態に識別した時の識別スコアを出力する。近年、一般物体認識の分野ではベクトル量子化された局所特徴のヒストグラムを用いて画像を認識する “bag-of-visual-words” という手法が注目されている¹³⁾¹⁴⁾¹⁵⁾。提案手法は、顔識別器ベースの bag-of-visual-words と呼べる手法であり、本論文で用いるサポートベクターマシンの他にも probabilistic Latent Semantic Analysis (pLSA) や Latent Dirichlet Allocation (LDA) 等の bag-of-visual-words の識別手法が適用可能と考えられる。

3. 顔検出手法

顔の向きと表情を抽出するために、向き毎（左向き、正面、右向き）の顔検出器と表情認識器を用いる。これらは、入力画像シーケンスの各フレームに適用する。顔検出手法を図 3 に示す。顔検出には、Viola と Jones により提案された Harr-like 特徴のカスケード接続による顔検出器 (VJ-detector) と、SVM に基づく非顔フィルターを組み合わせで用いている。同様な手法として、VJ-detector の最終段階で AdaBoost で選別した特徴をカーネル SVM で識別する手法も提案されている⁹⁾ が、ここでは、非顔フィルターでは、VJ-detector と異なる特徴量を用いることによって、False Positive (誤検出) を減らしている。これは、服や旗などに現れる勾配に着目すると顔らしいパターンは、輝度に着目すると顔ではないことが容易に解る場合があるという観察に基づく。VJ-detector により誤検出した領域の例を図 4

4 複数人物の表情認識による観客の満足度の自動定量評価



図 4 VJ-detector による誤検出領域の例

Fig. 4 Example of False Positive by VJ-detector.

に示す。非顔フィルターの画像特徴量と SVM のカーネルは表情認識手法と同じであり、それらは 4 章で説明する。異なる向きの顔が近い場所で検出された場合、非顔フィルターの識別スコアが最も高い向きの顔を採用する。この選択の実装は次の通りである。まず全方向の顔を検出し、識別スコアと顔の位置と大きさをスタックする。次に、識別スコアが大きい順に顔を選び、前に選んだ顔の矩形領域内に新しく選ぶ顔の中心がない時に顔とする。また、横向きの顔は、正面の顔を半分含む場所で検出されることが多いため、そのような位置ずれした顔も非顔サンプルとして学習して除去する。

4. 表情認識手法

4.1 局所特徴量

顔画像を 2 種類の表情（笑顔、非笑顔）に識別する識別器を顔の向き毎に用意する。現在研究されている表情認識手法は、目の中心や鼻の頂点等から数十の特徴的な部位を抽出して用いる手法¹⁸⁾とパターンベースで認識する手法⁶⁾とに大別されるが、本研究では、単純に実装することの出来るパターンベースの手法を用いる。図 5 に表情認識手法を示す。画像の特徴量として、5x5pixel 内の輝度ヒストグラム特徴 (Histogram of Intensity: HI) を用いる。これは、VJ-detector で検出された顔は位置が完全に合っておらず、小さな位置変動に頑健性を持たせるためである。特徴抽出の手順は次の通りである。まず、VJ-detector で検出された顔領域を 50x50pixel にリサイズし、ヒストグラム均一化を施す。次に、10x10 の重複のないセルに分割し、セル内でヒストグラムを計算する。ここで輝度値は、0~255 を 8 階調に分割する。HI 特徴の次元数は $10 \times 10 (\text{cells}) \times 8 (\text{intensity level}) = 800$ 次元である。特徴量に正規化は、セル毎にセル内のピクセル値の個数で割ることで行った。HI 特徴は極めて単純な特徴であるが、表情を認識するために十分な次元数を持っており、多数のサンプルを用いることで高い表情認識性能を持たせることが出来る。

近年、HOG(Histogram of Oriented Gradient) 特徴が注目されている¹⁰⁾。HOG は、ブロック内の勾配ヒストグラムであるため、HI 特徴のような明度変動の影響を受にくい。また Hu らは HOG を局所性保存射影 (Locality Preserving Projection:LPP) により次元圧

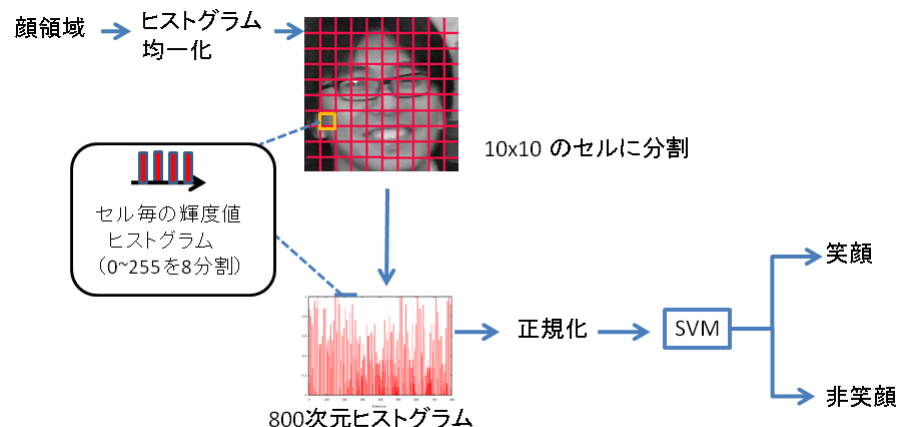


図 5 表情認識手法

Fig. 5 Facial Expression Recognition Method.

縮し識別性能を高めている⁷⁾が、HOG や LPP は、勾配のピンの切り方、正規化ブロックの取り方、正規化方法、圧縮する次元数等の設定パラメータが多い反面、HI は、設定パラメータが少なく、上記設定で十分な識別性能を持っていた。

4.2 サポートベクターマシン

識別器には、サポートベクターマシン (SVM)¹¹⁾を用いる。SVM は、識別面と学習サンプルとの最小距離 (マージン) を最大化するように識別面を決定する。線形 SVM の識別関数は以下のように表わされる。

$$f(\mathbf{H}) = \text{sgn}\left(\sum_{i \in SV} \alpha_i y_i \mathbf{H}_i \mathbf{H} - b\right),$$

ここで、 SV はサポートベクターの集合、 $\mathbf{H}_i, \mathbf{H}$ は学習サンプルとテストサンプル、 y_i はサンプル $\mathbf{H}_i$ のクラスラベル (+1 or -1)、 α_i は学習サンプル $\mathbf{H}_i$ に対する重み、 b は学習された閾値を表す。

ここまでは、線形分離可能な場合を仮定していた。線形分離不可能な場合には、非線形写像 $\psi(\mathbf{x})$ を用いることにより高次元空間にデータを射影し、高次元空間上でマージンを最大化することで対処できる。高次元空間での内積 $\psi(\mathbf{x})^T \psi(\mathbf{y})$ が、カーネル $K(\mathbf{x}, \mathbf{y})$ で計算可

能な場合、高次元空間に射影することなく学習や識別が行える。これは、カーネルトリックと呼ばれる。カーネル SVM の識別関数は以下で表わされる。

$$f(\mathbf{H}) = \text{sgn}\left(\sum_{i \in SV} \alpha_i y_i K(\mathbf{H}_i, \mathbf{H}) - b\right)$$

ここで、 $K(\mathbf{H}_i, \mathbf{H})$ は学習サンプル $\mathbf{H}_i$ とテストサンプル $\mathbf{H}$ のカーネル関数値を表す。本論文では、ヒストグラム特徴を用いて、表情を認識するために χ^2 カーネル¹²⁾ を用いる。ヒストグラム $\mathbf{H}_X = \{H_X(1), \dots, H_X(N)\}$ とヒストグラム $\mathbf{H}_Y = \{H_Y(1), \dots, H_Y(N)\}$ に対する χ^2 カーネル $K_{\chi^2}(\mathbf{H}_X, \mathbf{H}_Y)$ は以下のように定義される：

$$K_{\chi^2}(\mathbf{H}_X, \mathbf{H}_Y) = \exp\left(-\frac{1}{2\sigma} \sum_{i=1}^N \frac{(H_X(i) - H_Y(i))^2}{H_X(i) + H_Y(i)}\right).$$

カーネルパラメータ σ と SVM の正則化パラメータ $C^{(11)}$ は 5 分割交差確認法に基づくグリッドサーチによって決定する。

5. 満足度推定手法

5.1 投票方法

3 章と 4 章で述べた顔検出と表情認識は、入力動画の各フレームに対して適用されるものであった。顔向きと表情の認識結果は、1 つのビデオシーンスに対してヒストグラムに投票する。本論文では、顔表情 $\mathbf{H}^E$ (笑顔、非笑顔の 2 次元)、顔向き $\mathbf{H}^D$ (左向き、正面、右向きの 3 次元) と顔表情と顔向きの共起 $\mathbf{H}^{D \wedge E}$ (左向き $\wedge$ 笑顔, 左向き $\wedge$ 非笑顔, 正面 $\wedge$ 笑顔, 正面 $\wedge$ 非笑顔, 右向き $\wedge$ 笑顔, 右向き $\wedge$ 非笑顔の 6 次元) の 3 つのヒストグラムを用いる。各ヒストグラムはノルムが 1 になるように顔の個数で正規化する。ここで、顔の向きは観客の視線の向きの集中度を表現するために用いている。カメラの設定がシーンによって異なるため、顔の向きのヒストグラム $\mathbf{H}^D$ の次元の並びを以下のようにデータによって置き換えを行う。左向き顔の数が最も多いシーンでは、顔向きの次元の順序は (右向き, 左向き, 正面) とする。正面顔の数が最も多いシーンでは、顔向きの次元の順序は (左向き, 正面, 右向き) とする。右向き顔の顔が最も多いシーンでは、顔向きの次元の順序は (正面, 右向き, 左向き) とする。共起ヒストグラム $\mathbf{H}^{D \wedge E}$ の順序も同様に並び替える。これらは、フレーム毎に行う。この並び替えの効果は 7.3 で示す。共起ヒストグラムを用いたのは、試合を見ている人が笑っているのか、見ていない人が笑っているのかといった詳細な特徴をとるためである。以上の投票方法を識別器に対応させると図 6 のような流れになっている。

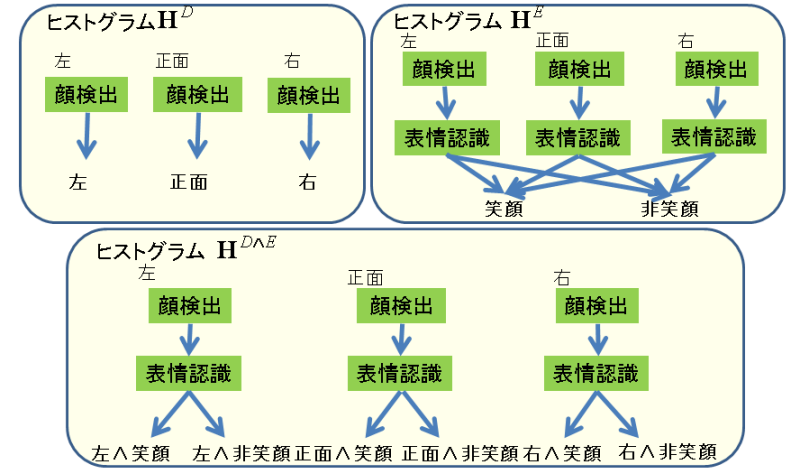


図 6 投票方法

Fig.6 Flow of Voting Method

5.2 サポートベクターマシンによる判別

映像中の観客の状態を識別するためにマルチチャンネル和カーネルを用いた SVM を用いる。識別対象に応じて、各ヒストグラムの重要性は異なると期待される。例えば、顔の表情は“喜んでいる”と“喜んでいない”を識別する際に重要であると考えられ、反対に顔の方向は観客が“真剣に観戦している”か“真剣に観戦していない”かを識別する際に重要であると考えられる。マルチチャンネル和カーネルにおいて各ヒストグラムの重要性を調整するために、各ヒストグラムの重みを以下のように導入する：

$$K_B(\mathbf{H}_X, \mathbf{H}_Y) = \alpha_D K_A(\mathbf{H}_X^D, \mathbf{H}_Y^D) + (1 - \alpha_D) K_A(\mathbf{H}_X^E, \mathbf{H}_Y^E) \\ \text{where } 0 \leq \alpha_D \leq 1.$$

ここで $K_A(\mathbf{H}_X, \mathbf{H}_Y)$ は以下で示すヒストグラムインターセクションカーネルである。

$$K_A(\mathbf{H}_X, \mathbf{H}_Y) = \sum_{i=1}^N \min(H_X(i), H_Y(i)),$$

ここで N は各ヒストグラムの次元数を示す。

6 複数人物の表情認識による観客の満足度の自動定量評価

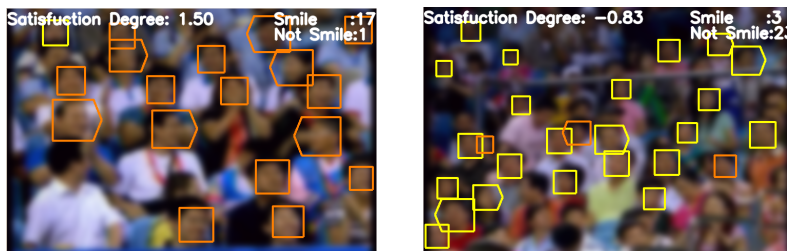


図 7 システム画面
Fig. 7 Screen of System

3つのヒストグラムを用いたマルチチャンネルカーネルを以下に示す。

$$K_C(\mathbf{H}_X, \mathbf{H}_Y) = \alpha_D K_A(\mathbf{H}_X^D, \mathbf{H}_Y^D) + \alpha_E K_A(\mathbf{H}_X^E, \mathbf{H}_Y^E) + (1 - (\alpha_D + \alpha_E)) K_A(\mathbf{H}_X^{D \wedge E}, \mathbf{H}_Y^{D \wedge E})$$

where $0 \leq \alpha_D + \alpha_E \leq 1$.

ヒストグラムインターセクションカーネルは Mercer カーネルである¹⁴⁾。Mercer カーネルの非負重み付き和カーネルも Mercer カーネルである¹⁶⁾ので、 K_B と K_C も Mercer カーネルである。

各ヒストグラムの重み α_D と α_E は交差確認法により自動的に決定することにより、各ヒストグラムの重要度に対するヒューリスティックは不要となる。

5.3 満足度推定

本論文では、“喜んでいる”か“喜んでいない”かを SVM により識別したときに“喜んでいる”状態に近い割合で満足度を表す。すなわち、“喜んでいる”データをクラスラベル 1、“喜んでいない”をクラスラベル-1 として学習した SVM を用いて、入力データ $\mathbf{H}$ の満足度は、満足度 ($\mathbf{H}$) = $\sum_{i \in SV} \alpha_i y_i K(\mathbf{H}_i, \mathbf{H}) - b$ とする。同様に“真剣に観戦している”か“真剣に観戦していない”かを SVM により識別したときのスコアを真剣度とする。

6 システム

3~5 章で述べた識別手法を linux-2.6 に実装した。システムは入力動画象に対して、検出結果、表情認識結果、満足度を表示するものである。VJ-detector には OpenCV^{*1}に実装

表 1 SVM の識別率 (学習データの 5 分割交差確認法による平均)。

Table 1 Classification performance of SVM(average of 5-fold cross validation on training data).

Direction	NonFace Filter	Expression(HI)
Front	95.62%	92.3%
Side	94.05%	93.5%

済みのものを利用した。VJ-detector の局所特徴量として、拡張 Harr-like feature⁴⁾を用いた。SVM の実装には LIBSVM^{*2}を用いた。左顔と右顔は対象であるため、左顔の識別器を全識別器 (VJ-detectors, 非顔フィルター, 表情認識器) で用意し、右顔の認識は入力画像を水平に反転させることで実現した。システム画面を図 7 に示す。四角形が正面顔を表し、五角形が左向きと右向きを表わしている。また、橙色が笑顔を表し、黄色が非笑顔を表している。

7. 実 験

7.1 識別器の学習

SVM の学習に用いたサンプルは 2 章で説明したデータセットと TV からキャプチャーした他の映像に VJ-detector を適用することで用意した。それらは、10frame 間隔の映像を用いている。顔の表情は手動により“笑顔”と“非笑顔”にラベル付けがされている。

表情認識において、正面顔は 6460 サンプル (笑顔 2730, 非笑顔 3730) と側面顔は 6440 サンプル (笑顔 2270, 非笑顔 4170) を用いた。正面顔の非顔フィルターには、3000 サンプルの顔画像と 3000 サンプルの非顔画像を用いた。側面顔の非顔フィルターには、2000 サンプルの顔画像と 2000 サンプルの非顔画像を用いた。ここで、VJ-detector で誤って検出された顔領域が非顔画像として用いられてる。各識別器の 5 分割交差確認法による識別率は表 1 に示す。

7.2 顔検出・表情認識実験

構築した識別器を用いた実際のシーンでの識別率を表 2 に示す。表 2 において、Accuracy は 検出された顔において、表情が正しく正解した割合を表す。横顔の表情に認識率は正面顔に比べて少し悪い。これは、横顔のバリエーションの多さによると考えられる。誤認識の主な原因は他の人間の頭や腕などによる隠れであった。隠れ領域が小さい場合には、大域的

*1 Software available at <http://www.sourceforge.net/projects/opencvlibrary>

*2 Software available at <http://www.csie.edu.tw/~cjlin/libsvm>

7 複数人物の表情認識による観客の満足度の自動定量評価

表 2 識別データに対する識別性能 FP(a): 間違った表情が混入する率, FP(b): 非顔が混入する率

Table 2 Accuracy on classification scene. FP(a): The incorporation rate of false expression, FP(b): The incorporation rate of non face

	Accuracy	FP(a)	FP(b)
Front(Smile)	96.3%	2.0%	1.7%
Front(Not Smile)	99.4%	0.3%	0.3%
Left(Smile)	93.8%	3.3%	2.9%
Left(Not Smile)	98.8%	0.2%	1.0%
Right(Smile)	89.7%	4.7%	5.6%
Right(Plane)	98.5%	0.2%	1.3%



図 8 隠れの存在する顔の例

Fig. 8 Example of Occluded Face.

カイ 2 乗カーネルより局所カイ 2 乗カーネルの和¹⁶⁾を用いることで改善が可能と考えられる。しかし、人間の眼でも識別が困難な大きな隠れを含む顔も存在していた。隠れの存在する顔の例を図 8 に示す。隠れに対応するためには、隠れているかを識別して、隠れの存在しない顔のみを用いることが必要と考えられる。これは今後の課題とする。隠れの場合を除いて、表情の識別率は高いといえる。次節では、観客状態の識別結果を述べる。

7.3 SVM による観客動画識別実験

観客の状態を線形 SVM とマルチチャンネル和カーネルを用いた SVM で識別した。誤識別率を leave-one-out 法で測定した。ベストパラメータにおける各手法の誤識別率を表 3 に示す。線形 SVM が、カーネル SVM に匹敵する性能が得られている。顔向きヒストグラムの順序を並び替えることによって、多くの場合識別率は 2~4% 改善した。“喜んでいる”が“喜んでいない”、“真剣に観戦している”が“真剣に観戦していない”かのどちらの識別タスクにおいても、重み付きヒストグラムインターセクションカーネル K_B, K_C は重みなしヒストグラムインターセクション K_A より 2~3% 識別率が改善した。3 チャンネルヒストグラム ($H^D, H^E, H^{D \wedge E}$) は 2 チャンネルヒストグラム (H^D, H^E) と同等以上の識別率であった。

線形 SVM で 3 チャンネルヒストグラム ($H^D, H^E, H^{D \wedge E}$) を用いた場合の各次元の重みを図 9 に示す。“喜んでいる”が“喜んでいない”かの識別においては、向きの並び替えを

表 3 シーン識別のエラー率 (%), (a): “喜んでいる” vs “喜んでいない”, (b): “真剣に観戦している” vs “真剣に観戦していない”, []: 向きの並び替えを行わない場合のエラー率。

Table 3 Error rates of scene classification (%), (a): “Positive Scene” vs “Negative Scene”, (b): “Watching Seriously” vs “Watching Not Seriously”, []: error rate of non aligned direction.

kernel	histogram	(a)	(b)
Linear	H^D	53 [44]	28 [40]
Linear	H^E	13 [-]	68 [-]
Linear	$H^{D \wedge E}$	10 [13]	32 [32]
Linear	H^D, H^E	13 [14]	30 [40]
Linear	$H^D, H^E, H^{D \wedge E}$	10 [13]	30 [32]
K_A	H^D	40 [39]	32 [36]
K_A	H^E	10 [-]	40 [-]
K_A	$H^{D \wedge E}$	13 [16]	34 [30]
K_A	H^D, H^E	12 [12]	32 [36]
K_A	$H^D, H^E, H^{D \wedge E}$	12 [13]	32 [34]
K_B	H^D, H^E	9 [12]	30 [34]
K_C	$H^D, H^E, H^{D \wedge E}$	9 [11]	28 [30]

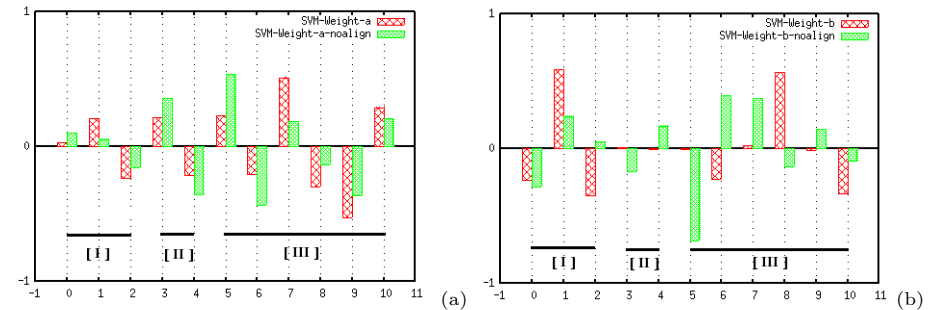


図 9 線形 SVM の $H^D, H^E, H^{D \wedge E}$ に対する重み。[I]: H^D - 左から順に 左、正面、右, [II]: H^E - 左から順に 笑顔、非笑顔 [III]: $H^{D \wedge E}$ - 左から順に 左 ∧ 笑顔、左 ∧ 非笑顔、正面 ∧ 笑顔、正面 ∧ 非笑顔、右 ∧ 笑顔、右 ∧ 非笑顔、赤棒: 向きの並び替えあり、緑棒: 向きの並び替えなし, (a): “喜んでいる” vs “喜んでいない”, (b): “真剣に観戦している” vs “真剣に観戦していない”。

Fig. 9 Weights of linear SVM for each dimension of histograms, $H^D, H^E, H^{D \wedge E}$. [I]: H^D - the order is left, front, right, [II]: H^E - the order is smile, not smile [III]: $H^{D \wedge E}$ - the order is left ∧ smile, left ∧ not smile, front ∧ smile, front ∧ not smile, right ∧ smile, right ∧ not smile, red bar: aligned direction, green bar: non aligned direction, (a): “Positive Scene” vs “Negative Scene”, (b): “Watching Seriously” vs “Watching Not Seriously”。

行うことによって、最も多く向いている方向の人が、笑顔であることを示す次元が“喜んでいる”を最もよく支持している。一方で並び替えを行っていないときには、左向きの顔が笑顔であることを示す次元が高い重みとなっている。また、“真剣に観戦している”と“真剣に観戦していない”においても向きの並び替えを行うことによって、最も多く向いている方向を示す次元と最も多く向いている方向の顔が非笑顔であることを示す次元が“真剣に観戦している”を最もよく支持している。このように最も向いている方向の向きを揃えるように並び替えを行うことによって適切な重みを獲得することが出来、識別率が向上したと考えられる。

SVM の同じ C パラメータにおいて α_D に対応したエラー曲線を図 10 に示す。図 10 においてパラメータ α_D は 0.05 間隔で変化させている。“喜んでいる”が“喜んでいない”かを識別する場合、最適なパラメータは $\alpha_D = 0.25$ であり、誤識別率は α_D を大きくするに従い増加している。これは、この識別問題に対して顔の表情の重要性が高いことを意味している。反対に“真剣に観戦している”が“真剣に観戦していない”かの識別の場合、最適な重みは $\alpha_D = 0.55$ であり、 α_D を小さくするに従って誤識別率は増大している。これは、このタスクにおいて顔の向きがより重要であることを意味している。これらの重要性は、我々の予測と一致した。それは、表 3 において、ヒストグラム H^D を用いた場合には、“真剣に観戦している”が“真剣に観戦していない”かの識別率が高く、 H^E を用いた場合には、“喜んでいる”が“喜んでいない”かの識別率が高いことから解る。また、カーネル K_C の最適重みは、“喜んでいる”が“喜んでいない”かを識別する場合、 $\alpha_D = 0, \alpha_E = 0.55$ であり、“真剣に観戦している”が“真剣に観戦していない”かを識別する場合、 $\alpha_D = 0.4, \alpha_E = 0.45$ であった。

誤識別したデータは、“喜んでいる”状態で観客は騒いでいるが笑っていない、“真剣に観戦している”シーンで異なる方向を見ているなど我々のアプローチから外れたシーンで起こっていた。騒いでいる状態を識別するためには、叫びの表情を追加したり、手を振るなど動きの情報を追加することが必要となる。

7.4 満足度推定結果に関する考察

次に満足度について、最適パラメータを用いた場合に対する考察を行う。図 11 は線形 SVM とカーネル K_C における満足度の分布である。これらは、7.3 節の識別タスク (leave-one-out) におけるテストデータの分布である。“喜んでいる”が“喜んでいない”の識別スコア (満足度) は、“真剣に観戦している”と“真剣に観戦していない”の識別スコア (真剣度) に比べ、正解ラベルの方によく分離して分布している。しかし、図 11 a のように飛

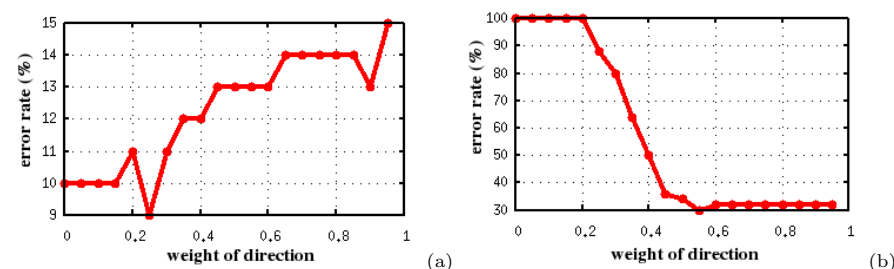


図 10 向きの重み α_D に対応するカーネル K_B のエラー率 (leave-one-out). (a): “喜んでいる” vs “喜んでいない” ($C=32$), (b): “真剣に観戦している” vs “真剣に観戦していない” ($C=0.5$).

Fig. 10 Error rates for K_B corresponding to direction weight α_D (leave-one-out). (a): “Positive Scene” vs “Negative Scene” ($C=32$), (b): “Watching Seriously” vs “Watching Not Seriously” ($C=0.5$).

び跳ねたり、拍手したり、叫んだりして喜んでいるが笑ってはいない人が多いデータに対しては、低い満足度が与えられるなど、正解が“喜んでいる”データが“喜んでいない”と判定されるようなデータに対し、不自然な満足度となっていた。“喜んでいる”が“喜んでいない”を識別する場合 (L-a),(KC-a) において図 11 b のように満足度が低いと思われるシーンでは満足度が低く図 11 c,d のように、笑っている人の数が中程度であれば、中程度の満足度となっており図 11 e のように全員笑っているようなシーンでは高い満足度となった。正解が“喜んでいない”データに対しては、図 11 f のように楽しそうでない場合に満足度が低く、図 11 g のように楽しそうになるに従って、満足度は高くなっていき図 11 h,i,j のように楽しそうにしている人が多いと、満足度が高くなった。これは、試合前に楽しそうにしている観客のデータが“喜んでいない”サンプルに入っているからであり、これらが満足していると判定されるのは、自然であると考えられる。

次に“真剣に観戦している”と“真剣に観戦していない”の識別タスク (L-b)(KC-b) において考察する。図 11 k においては、一部の人は真剣に観戦しているデータだが、色々な方向を向いている人がいるため真剣でないとして判定されてしまっている。図 11 l,n においては見た目の真剣度が上がるに従い真剣度が上がっている。“真剣に観戦していない”データに関しても図 11 m は、全く真剣に観戦していないと思われるデータであり、真剣でないとして正しく判定されている。図 11 o においては、手を振ったり、横を向いたり真剣に観戦していない人も真剣に観戦している人もいて、中間的な真剣度と考えられるデータであったが、正しく真剣度が推定されている。図 11 p は正解が“真剣に観戦していない”データである

9 複数人物の表情認識による観客の満足度の自動定量評価

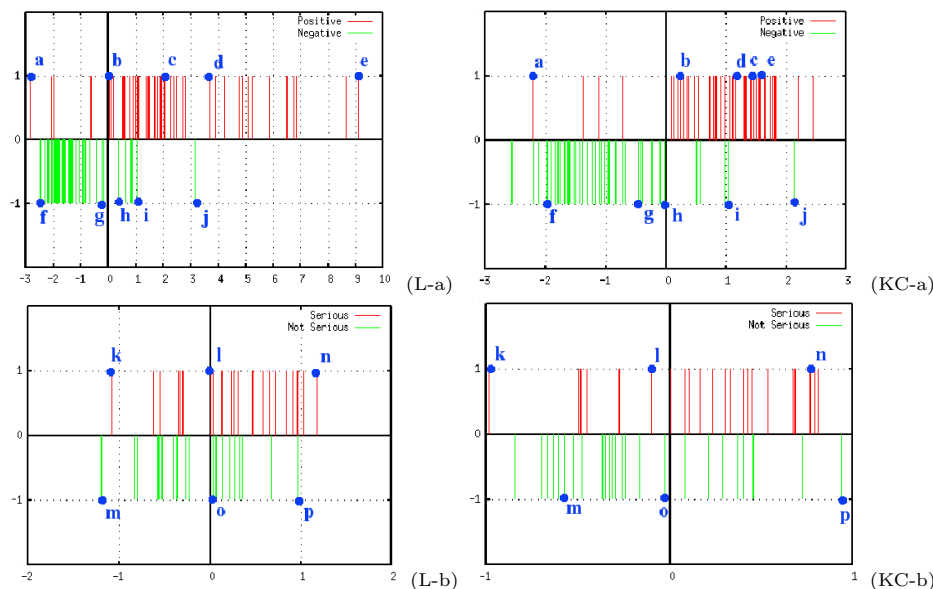


図 11 満足度の分布: 横軸が識別スコアを表し、縦軸が正解クラスラベルを表す。(L-a): 線形 SVM、“喜んでい
る”vs“喜んでいない”, (KC-a): カーネル K_C 、“喜んでい
る” vs “喜んでいない”(L-b): 線形 SVM、
“真剣に観戦している”vs“真剣に観戦していない”, (KC-a): カーネル K_C 、“真剣に観戦” vs “真剣に観
戦していない”, ラベル a~p のデータ例の説明表 4 に示されている。表 4 には、各データの説明とヒストグ
ラム H^E , H^D の数値が示されている ($H^{D \wedge E}$ は省略)。

Fig. 11 Distribution of the degree of spectators' satisfaction: horizontal axis indicates clasifica
tion score, vertical axis indicates correct class label. (L-a): Linear SVM,“Positive Scene ”
vs“Negative Scene ”, (KC-a): Kernel K_C ,“Positive Scene” vs “Negative Scene ”(L-b): Linear SVM,“Watching Seriously
ly”vs“Watching Not Seriously”, (KC-a): Kernel K_C ,“Watching Seriously” vs “Watching Not Seriously”, The explanation of the data of label a~p are shown
in Table.4. In Table.4, The explanation of each data and histogram H^E , H^D are shown.
($H^{D \wedge E}$ are abbreviated).

が、真剣に観戦していない人もいるが、真剣に観戦している人も多かったため、真剣度が高
まったと考えられる。このように、“真剣に観戦している”と“真剣に観戦していない”の
識別率は、低くなっているがそれはラベル付けを試合前が試合中に優勢でない状況かで、“
真剣に観戦している”か“真剣に観戦していない”かをラベル付けしたためであり、“真剣
に観戦している”か“真剣に観戦していない”でないかという意味ではミスラベルが混入し

表 4 図 11 におけるデータの説明
Table 4 Explanation of data in Fig.11

label	観客の状態	笑顔	非笑顔	左	正面	右
a	拍手して喜んでいる	0.03	0.97	0.07	0.91	0.02
b	静かに喜んでいる	0.02	0.80	0.03	0.84	0.13
c	様々な方向を向いている	0.38	0.62	0.38	0.47	0.15
d	多くの人が喜んでいる	0.57	0.43	0.23	0.76	0.01
e	全員喜んでいる	0.90	0.10	0	0.99	0.01
f	楽しそうではない	0	1	0.46	0.54	0
g	一部の人は拍手している	0.12	0.8	0.04	0.54	0
h	棒を叩いて応援している	0.23	0.77	0.06	0.68	0.26
i	観客同士話をしている	0.27	0.73	0.36	0.49	0.15
j	一部楽しそうにしている	0.58	0.42	0	0.71	0.29
k	半分ぐらいの人は真剣	0.08	0.92	0.12	0.47	0.41
l	多くの人が真剣に観戦している	0.14	0.86	0.31	0.69	0
n	全員は真剣	0.01	0.99	0	1	0
m	真剣でない	0	1	0.15	0.43	0.42
o	一部の人は真剣	0.19	0.81	0.07	0.70	0.23
p	一部の人は真剣	0	1	0	0.94	0.05

ていると考えられる。よって、性能を正しく評価するためには、正しいラベル付けをする必
要がある。

8. ま と め

本論文では、映像中の顔に着目して、観客の満足度を推定するシステムを提案した。提案
システムは、顔識別器ベースの bag-of-visual-words に基づいており、顔表情の識別器を画
像上に検出された全ての顔に適用することにより複数の人物の表情を認識し、表情と顔の向
きを投票したヒストグラムを作る。このヒストグラムを特徴ベクトルとした識別器を用い
て観客の満足度を自動的に推定する。提案システムを用いて TV から収集した観客の映像
を用いた識別実験を行い、“喜んでい
る”と“喜んでいない”または“真剣に観戦している”
と“真剣に観戦していない”場面をある程度識別できることを示し、その識別スコアに基づ
いて満足度を推定した。“喜んでい
る”と“喜んでいない”を識別する場合には顔表情が重
要であり、“真剣に観戦している”と“真剣に観戦していない”を識別する場合には顔向きが
重要であるということも確認できた。

今後は顔表情と向きのクラスを増し、より多くの観客の状況を識別すること、教師無し学
習で観客の状況を分類するなどより適切な観客のラベル付け法を検討することが課題とし

てあげられる。

参 考 文 献

- 1) P.Viola and M.J.Jones, Robust Real Time Face Detection , IJCV 57(2), pp.147-154, 2004.
- 2) C.Haung, H.Ai, Y.Li, S.Lao, High-Performance Rotation Invariant Multiview Face Detection, IEEE Trans. on PAMI, Vol.29, No.4, pp.671-686, 2007.
- 3) B.Wu, H.Ai, C.Haung and S.Lao, Fast Rotation Invariant Multi-View Face Detection Based on Real Adaboost, In FG, 2004.
- 4) R. Lienhart and J. Maydt, An Extended Set of Haar-like Features for Rapid Object Detection, In ICIP, Vol. 1, pp. 900-903, Sep. 2002.
- 5) A.Hidaka and T.Kurita, Non-Neighboring Rectangular Feature Selection Using Particle Swarm Optimization, In ICPR, 2008.
- 6) Y.Shinohara, N.Otsu, Facial Expression Recognition Using Fisher Weight Maps, In FG, 2004.
- 7) Y.Hu, Z.Zeng, L.Yin, X.Wei, X.Zhou and T.S.Haung, Multi-View Facial Expression Recognition, In FG, 2008.
- 8) H-Y.Chen, C-L.Haung and C-M.Fu, Hybrid-boost learning for multi-pose face detection and facial expression recognition, Pattern Recognition, 41, pp.1173-1185, 2008.
- 9) D.D.Le and S.Satou, Feature Selection By AdaBoost For SVM-Based Face Detection, FIT 2004, pp.183-186, 2004.
- 10) N.Dalal and B.Triggs, Histograms of Oriented Gradients for Human Detection, In CVPR, pp.886-893, 2005.
- 11) V.Vapnik, Statistical Learning Theory, John Wiley & Sones, 1998.
- 12) O.Chapelle, P.Haffner and V.Vapnik, SVMs for Histogram Based Image Classification, IEEE Trans. on NN, Volume 10, Issue 5, pp.1055-1064, 1999.
- 13) S.Lazebnik, C.Schmid, and J.Poncet, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, In CVPR, pp.2169-2178, 2006.
- 14) K.Grauman and T.Darrell, The Pyramid Match Kernel: Discriminative Classification with Set of Image Features, In ICCV, pp.1458-1465, 2005.
- 15) T.Matsukawa, K.Suzuki and T.Kurita, Preliminary Local Features Selection by Support Vector Machine for Bag of Features, In 15th Japan-Korea Joint Workshop on Frontiers of Computer Vision (FCV), 2009.
- 16) K.Hotta, Robust face recognition under partial occlusion based on support vector machine with local Gaussian summation kernel, Image and Vision Computing, Vol.26, Issue 11, 1, pp.1490-1498, 2008.
- 17) 山本 誠, 谷本弘昭, 新田直子, 馬場口登, 個人的選好獲得のための特定人物のテレビ

視聴時における興味区間推定, 電子情報通信学会論文誌, Vol.J90-D-II, No.8, pp.2202-2211, 2007.

- 18) 宮原正典, 青木政樹, 滝口哲也, 有木康雄, 顔表情からの関心度推定に基づく映像コンテンツへのタギング, 情報処理学会論文誌, Vol.49, No.10, pp.3694-3702, 2008.