1V-05

集団活動時の楽しい振り返りを支援する身体装着型カメラによる体験自動記録

木下恵理子† 小坂真美†† 藤波香織†††
東京農工大学 工学部 情報工学科† 東京農工大学 工学府 情報工学専攻††
東京農工大学 大学院 工学研究院 先端情報科学部門†††

1. はじめに

近年，カメラの小型化や省電力化により，GoPro[1]やNarrative clip[2]などのウェアラブル型カメラが普及している．これらのデバイスを用いて自動的または無意識的に記録を行うライフログへの関心が高まっている．それに伴い，様々なセンサ情報から状況を推定し，タグ付けや分類を行う研究がなされている[3][4]．しかし，分類結果に対する利用者の満足度に言及する研究は少ない．また，複数のセンサ搭載により利用時に抵抗感や不自然さが生じると考えられる．そこで本研究では，映像や音の情報のみを用いて，利用者が思い出を振り返る際に楽しさを感じる場面の自動判定手法を開発する．

2. 場面自動判定システム

2.1 前提

2015年6月に「楽しさを感じる場面」についてのアンケートを行い，自動撮影に関する意識調査を行った．その結果，「きれいな風景」や「珍しいもの」の自動撮影は意味がないという回答が多く，「誰かと会話している様子」や「盛り上がった状態」の自動撮影に高い需要があることが分かった．よって，「会話風景」「興味」「盛り上がり」という 3 つの場面に焦点を当てた．「興味」は，自動撮影によりユーザが無意識に注目したものを思い出として残すことが可能であると考え，追加したものである．

(A) 会話風景：会話の中で，ユーザが特に思い出深い区間
(B) 興味　　：撮影者が何かを確認・注視した区間
(C) 盛り上がり：笑いや大声が上がっている区間

これら 3 場面について，ユーザが思い出を振り返る際に楽しさを感じる場面を抽出する．

2.2 システム概要

本研究では映像と音の情報のみで判定するため，ウェアラブルカメラで予め撮影した動画ファイルを入力とする．また，データ容量や振り返り時の手軽さを考慮し，静止画を出力とする．

処理の流れを図 1 に示す．動画から得られる映像データと音声データから，「会話風景」「興味」「盛り上がり」の判定に必要な特徴量を一定区間ごとに算出し，3 つの分類器を用いて機械学習での場面推定を行う．このとき推定結果は，分類結果の確からしさを示す0～1の値である．その後，推定結果を単純加算したものを「撮影度」と定義し，撮影度の高い区間のフレームを保存する．「会話風景」は笑顔検出や，会話中の相手への集中具合などを見る注視度から検出する．「興味」は注視度から，「盛り上がり」は音量や笑い声から検出を行う．

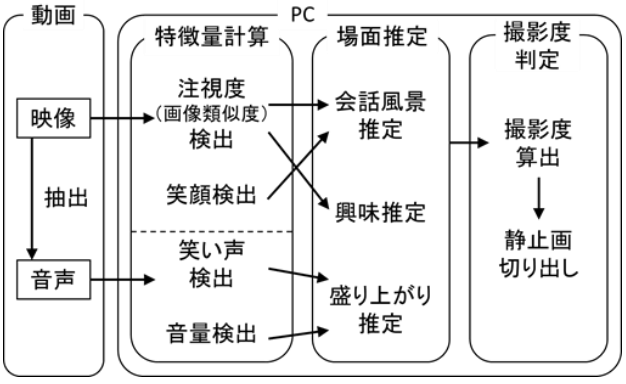

図 1　システム概要

なお，本稿執筆時点では笑顔検出部が未実装であり，関連のある「会話風景」を除く 2 場面の推定について実装を行い，撮影度は 2 場面の推定結果の単純加算とした．また，撮影度は 1 秒単位で算出されるが，保存するフレームはブレの大きさなどを考慮し選択するものとする．

2.3 判定手法

2.3.1 特徴量の検討と計算手法

本手法では，映像から「注視度」「笑顔」，音声から「笑い声」「音量」の計 4 項目を特徴量として算出し，3 場面を推定する．

(i) 注視度

注視度に関する特徴量は，0.25 秒前および 0.5 秒前のフレームで算出したヒストグラムとの差分値の 1 秒ごとの和である．また，得られた注視度と 1 秒前の注視度の差も特徴量とする．前フレームとの注視度は，値が小さいほど大きく，値が大きいほど小さくなる．8 ビットの RGB 画像に対して，グレースケール化と減色処理を行い，8 色，64 色，256 色へと変換した後にそれぞれのヒストグラムを計算し，12 次元の特徴量とした．

(ii) 笑い声

笑い声に関する特徴量は，音声データから基本周波数を算出した際の 1 秒間での上昇値とする．基本周波数は，式(1)で定義される自己相関関数において，R(j)が大きい値をとる上位 3 つのうち j が最小となるものを用いた．

$$\mathrm{R(j)} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N} v(i)\cdot v(i+j) \quad j = 0, 1, 2, \ldots, N \quad (1)$$

周期 j の逆数である基本周波数は，0.5 秒幅，0.25 秒幅，0.1 秒幅でそれぞれ算出し，1 秒間での上昇量の和とした．

また，1 秒ごとの隣接した複数の周波数成分の最小値を推定ノイズとし，周波数成分から動的に差し引いたデータや，ノイズと考えられる小振幅の箇所を縮小したデータも同様に基本周波数を計算し，12 次元の特徴量とした．

(iii) 音量

音量に関する特徴量は，音声データの波形から一定区間（0.1 秒，0.5 秒，1 秒）ごとに最大値をとり，値そのものや，1 秒前の値との差を取った値の 12 次元を用いる．

2.3.2 分類器の作成

前節で定義した計 36 次元の特徴量を用いて，Weka[5]を用いた分類器の作成および評価を行う．分類は「興味」「盛り上がり」それぞれに対して行い，分類手法は F 値の高い結果が得られた RandomForest（決定木数 100）とした．また，それぞれの分類器において最も F 値が高くなるような特徴量選択の結果における貢献度は，「興味」の分類器では 1 秒前の注視度との差をとった特徴量すべてが低く，8 色に減色して計算を行った特徴量が最も高かった．「盛り上がり」の分類器では，笑い声の特徴量が大幅に低く，音量の最大値そのものを用いた特徴量が高くなった．

作成した分類器を用いて，事前に収集したデータに対し 10 分割交差検定を行った結果を混合行列で表 1 に示す．F 値は「興味」で 0.731，「盛り上がり」で 0.810 となった．

Automatic Recording of Group Activities for Enjoyable Recall by a Wearable Camera

Eriko KINOSHITA†, Mami KOSAKA††, Kaori FUJINAMI†††
†, ††, †††Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology

表 1　10 分割交差検定結果（①：興味，②：盛り上がり）

	①	他
①	208	84
他	73	219

	②	他
②	270	54
他	69	255

3.　評価実験

3.1　実験概要

本システムにより得られた静止画がユーザに与える感情的作用および妥当性を評価する実験を，3 人ずつ 2 組，計 6 人の被験者で行った．実験は「撮影」「インタビュー」「切り出し」の順に行った．「撮影」では，表 2 に示した 3 つのイベントに参加する様子を 15～20 分ずつ撮影した．このとき撮影者（カメラ装着者）はイベントごとに交替する．「インタビュー」では，システム出力とランダム選択により動画から切り出された各 30 枚の静止画を順不同で見せ，静止画が被験者に与える印象を調査した．このとき，被験者は静止画に対し「自分の電子アルバムに残したいと思うか」の 5 段階評価と理由を回答する．「切り出し」では，撮影者（カメラ装着者）に動画を見せ，2.1 で述べた(A)～(C)の 3 場面に「(D) 自由：自分の観点で“この場面を残したい”と思った区間」を加えた 4 項目に関して 1 秒単位でのラベル付けをしてもらい，システムの出力と比較を行った．

表 2　イベントの性質

イベント	性質
カードゲーム	会話の相手を見ていない可能性が高い 視点移動が少なく，周期性がある
机を囲んでの談笑	会話の相手を見ている可能性が高い 視点移動が少なく，不規則である
散歩	会話の相手を見ていない可能性が高い 視点移動が多く，不規則である

3.2　実験結果

被験者は 6 人だが，「インタビュー」以降の実験において 1 人分の結果を得ることができなかったため，5 人分の結果のみ示す．

3.2.1　インタビュー

実験で撮影した 6 つの動画から得られた静止画群に対し，システムが算出した撮影度と被験者の 5 段階評価の結果を比較する．静止画は計 180 枚であるが，本稿では一部の結果を示す．

6 動画の撮影度と得点の相関係数を計算した結果，-0.050～0.313 となり，全体的に小さい値となった．

全員が高得点を付けた静止画に対しては「全員が笑っていて楽しそう」「状況がよくわかる」「全員がきれいに写っている」という意見が多く，最低点を付けた静止画に対しては「ブレが大きい」「誰も写っていない」「カメラの向きが悪い」「既に似た写真を見た」「状況が伝わらない」という意見が多かった．特にブレやカメラの向きに関しては，得点を左右する大きな要因となった．また，撮影度が低く得点が高い静止画に関しては，「珍しい」「意外性がある」という新たな意見が得られた．

3.2.2　切り出し

「興味」「盛り上がり」それぞれに対し，撮影者（カメラ装着者）が付けたラベルと撮影度を比較する．すべての動画のデータにおけるラベルの有無と撮影度の箱ひげ図を図 2 に示す．

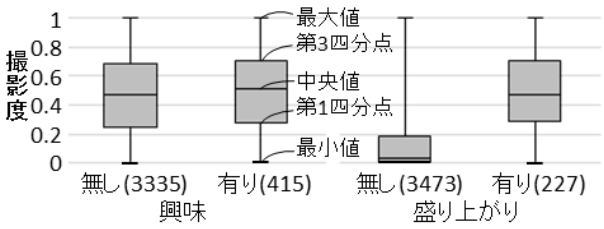

図 2　ラベルの有無と撮影度（括弧内はデータ数）

また，ラベルの有無と撮影度の関係を有意水準 5%で t 検定を行った結果，「興味」は有意差なし，「盛り上がり」は有意差ありという結果となった．また，動画ごとにデータを分け，同様にそれぞれ有意水準 5%で t 検定を行った．その結果，いずれの動画においても，すべての動画のデータでの結果と同様，「興味」では有意差なし，「盛り上がり」では有意差ありとなった．

また，(D)でユーザがラベル付けした区間は，以下の 4 場面に大別された．

(ア)　笑いが起きた場面
(イ)　全員が笑顔になった場面
(ウ)　珍しい顔が撮影された場面
(エ)　思わぬ失敗をして声が上がった場面

3.3　考察

3.3.1　インタビュー

撮影度とユーザ評価の相関が小さい原因として，各検出部での精度が低いことが考えられる．2.3.2 で述べたように「盛り上がり」での「笑い声」に関する特徴量の貢献度も低いため，計算方法を再考する必要がある．ブレの有無や写っている人数，カメラのアングルが得点に影響を与え，類似した静止画があると得点が低くなるといえる．また，珍しい，もしくは状況が伝わりやすいような場面が高得点となることが考えられる．

3.3.2　切り出し

3.2.2 から，システムの分類とユーザのラベル付けに関して，「興味」と比べると「盛り上がり」は適切な切り出しが行われていることが分かる．しかし，ラベル有りでの撮影度の中央値が 0.5 を下回っていることから，改善の余地があるといえる．「興味」の分類が適切でない原因として，特徴量の種類が不十分であることが考えられる．そのため今後は，注視度だけではなく新たな特徴量を考案する必要がある．「盛り上がり」の分類が適切でない原因としては，3.3.1 でも述べたように，各検出部での精度が低いことが考えられる．また，3.2.2 で挙げた 4 つの場面に関して，(ア)，(イ) は現在のアプローチで今後解決できる可能性が高いが，(ウ)，(エ) は別のアプローチが必要であると考えられる．

4.　おわりに

本稿では，映像や音の情報から利用者が思い出を振り返る際に楽しさを感じる場面の自動判定を行う手法を考案し，一部を実装したシステムでの評価を行った．「盛り上がり」についてはある程度成果が得られたが，本手法による場面判定にはいまだ改良の余地があるため，新たな特徴量や計算手法を挙げ，各検出部での精度を向上させる必要がある．

また，今後の課題として，以下が挙げられる．

- フレーム内の笑顔とその度合いを検出する機能の実装
- 出力結果から類似した静止画を除く機能の実装
- カメラのアングルを考慮した特徴量の考案
- 珍しい場面を検出する手法の考案
- 状況が伝わりやすい場面を検出する手法の考案
- ブレの少ないフレームを自動選択する機能の実装

謝辞

本研究の一部は科研費基盤(C) 15K00265 の支援を受けた．

参考文献（web ページはすべて 2016-01-05 閲覧）

[1]　Woodman Labs; “GoPro”, http://jp.gopro.com/
[2]　Narrative; “Narrative Clip 2 – The world’s most wearable camera” http://getnarrative.com/
[3]　堀鉄郎，相澤清晴；“ライフログビデオのためのコンテキスト推定”，電子情報通信学会技術研究報告，IE，画像工学，103.514，pp.67-72，2003
[4]　中村裕一；“映像によるライフログ”，情報の科学と技術，pp.57-62，2013
[5]　The University of Waikato “Weka”, http://www.cs.waikato.ac.nz/ml/weka/