7V-07

身体装着型カメラの映像を用いた集団活動時の楽しい振り返りのための体験自動記録 —会話場面の判定および評価—

木下恵理子† 藤波香織††
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，GoPro[1]や Narrative clip[2]などのウェアラブル型カメラを用いた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，画像や映像によるライフログにおけるセンサ情報を用いた場面推定や要約を行う研究が多くなされている[3][4]．一方，「思い出を振り返る楽しさ」に着目した研究は少なく，多くのセンサ利用は自然な記録の妨げとなる．そこで本研究では，一人称視点で撮影された映像や音の情報のみを用いて，撮影者や場を共有した人が楽しさを感じる場面の自動判定手法を開発する．本稿では，既報[5]で定義した「会話風景」「盛り上がり」「興味」の 3 場面のうち，未実装の「会話風景」の判定手法を実装し，オフライン評価を行う．

2. 会話風景判定手法

2.1 「会話風景」の定義

2015 年 6 月に行ったアンケートでは，画像や映像によるライフログにおいて，周囲の人との会話場面を思い出として残したいという意見が多かった．しかし，会話場面は日常生活において多く収集され，視覚的な情報の変化が小さい場面である．そのため，会話場面の中でより思い出深く感じることのできる区間を保存する必要がある．よって本研究では，「会話風景」を「会話の中でユーザが特に思い出深いと感じる区間」と定義し，ユーザが思い出を振り返る際に楽しさを感じる場面の検出を行う．

2.2 判定手法概要

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．

会話風景判定の処理の流れを図 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」の判定に必要な特徴量を 1 秒ごとに算出し，機械学習により構築した分類器を用いた場面推定を行う．学習に用いる特徴として，被写体であるユーザが笑顔であればあるほど思い出として残したい欲求が高まるという仮定から「笑顔」を用いた．また，写っている人数が多い静止画が好まれるという知見[5]から「人数」，会話内容の盛り上がりと思い出深さに相関があると仮定し，「盛り上がり」の場面判定[5]でも特徴として用いた「音量」を用いる．

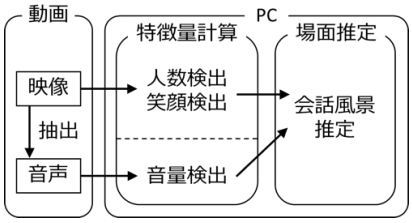

図 1 システム概要

Detection of Conversational Scenes towards Visual Lifelogging for Enjoyable Recall of Group Activities

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

2.3 特徴量の検討および計算手法

本手法では，映像から「人数」「笑顔」，音声から「音量」の計 3 項目を特徴量として算出し場面推定を行う．

(i) 人数

人数に関する特徴量は，1 秒間のフレーム群から等間隔で選択した 10 フレームにおいて顔検出を行った結果から算出する．顔検出は，Google 社が提供する機械学習の画像認識サービスである Cloud Vision API[6]の顔検知を用いている．10 フレームで検出された顔の数の平均値，最大値，最小値，分散値，およびこれら 4 特徴量の 1 秒前の値との差の計 8 次元の特徴量とした．

(ii) 笑顔

笑顔に関する特徴量は，「人数」と同様に 10 フレームでの顔検知結果から 1 秒分の特徴量を計算する．Cloud Vision API では笑顔の度合いを「VERY_LIKELY」「LIKELY」「POSSIBLE」「UNLIKELY」「VERY_UNLIKELY」の 5 レベルで表現しているため，表 1 のように対応させた．また，1 フレーム内で顔が複数検出された場合は複数のスコアが存在することになるため，スコアの平均値，最大値，最小値，合計値を計算する．その後，10 フレームで算出されたスコアの平均値の合計，合計値の合計，最大値，最小値，平均値の分散値，およびこれら 5 特徴の 1 秒前の値との差の計 10 次元の特徴量とした．

表 1 笑顔の度合いのスコア化

レベル	点数
VERY_LIKELY	1.00
LIKELY	0.75
POSSIBLE	0.50
UNLIKELY	0.25
VERY_UNLIKELY	0.00

(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．

3. 分類器の選択

2 節で定義した 30 次元の特徴量を用いて，機械学習ツール Weka[7]を用いた分類器の作成を行う．用いるデータは，平均撮影時間 24 分（min: 13 分, max; 48 分）の 10 動画である．

3.1 分類手法の選定

判定は，撮影者や周囲の人が実際に動画を閲覧し 1 秒単位でラベル付けした「会話風景」と，それ以外の区間を示す「その他」の 2 クラスとする．分類は，代表的な手法である NaiveBayes や Multilayer Perceptron，Sequential Minimal Optimization，J48 等と比較し，基本的な精度指標である全体の F 値が最も高い値 (0.852) となった RandomForest (決定木数 100) とした．また，「会話風景」の取りこぼしの最小化が重要であるため，F 値に加え「会話風景」の再現率を精度評価の指標とする．「その他」を「会話風景」と誤判定することは，意外性のある場面を

抽出する可能性があり，必ずしも悪影響ではないと考える．

3.2 データ不均衡の調整

性質上クラスのサンプル数に大きな偏りがあることから，前処理として Weka で実装されている SMOTE ef471828579 ¥r [8]によるオーバーサンプリングや Spread Subsample によるアンダーサンプリングを行い，学習データにおける不均衡を解消する[9][10]．10 分割交差検証でこれらのサンプル数増減方法を用いて精度を検証したところ，SMOTE で「会話風景」を「その他」と同数にオーバーサンプリングすることで「会話風景」の再現率 0.884，全体の F 値 0.919 と最も高い値となった．よって，前処理として SMOTE を用いたオーバーサンプリングを行う．

3.3 特徴量選択

10 分割交差検証において最も F 値が高くなる特徴量の組み合わせを選択した．相関に基づく属性補集合評価[11]と逐次貪欲探索による組み合わせ検証では，人数に関する特徴の貢献度が高かった．しかし，全特徴量を用いた場合が最も F 値が高かったため，以降の実験では全ての（30 次元）特徴量を用いる．

4. オフライン評価と考察

4.1 1動画抜き交差検証

10 動画から得られるデータセットに対して，9 動画を用いて学習をして残りの 1 動画でテストを行う「1 動画抜き交差検証」を行い，精度評価を行う．各動画の長さの違いによる結果への影響力の差を排除するため，最も長い動画に合わせサンプル数を正規化し平均をとった．その結果を混合行列で表したものが表 2 である．この結果から，「会話風景」の正解率が低く，「その他」を「会話風景」として多く分類していることが分かる．

表 2　1動画抜き交差検証（値は小数点以下 2 位を四捨五入）

		予測	
		会話風景	その他
正解	会話風景	16.5	331.3
	その他	177.5	2370.7

また，「会話風景」のうち，正解のデータの区間と不正解のデータの区間におけるフレームの一部を図 2 に示す．この結果から，正解と不正解のフレームに関して，視覚的に大きな違いは見られなかった．よって，本来は「会話風景」となるべき「その他」への誤分類についても，正解した区間との視覚的類似度や時間的な距離を考慮した判定を付加することで救済でき精度向上に繋がると考えられる．

図 2　「会話風景」フレームの一例（上：正解，下：不正解）

4.2 特徴ごとの分類精度と貢献度

2.3 節で述べた特徴ごとに「会話風景」の再現率と F 値を比較する．全ての特徴量を用いた場合と，「人数」「笑顔」の映像特徴のみを用いた場合，「音量」の音声特徴のみを用いた場合，また，同様に「人数」「笑顔」それぞれの特徴量のみを用いた場合の精度（「会話風景」の再現率および全体の F 値）を図 3 に示す．図から，音声の特徴のみを用いた場合の再現率が特に高いことが分かる．このことから，ユーザが会話風景として残したい場面は，会話自体が盛り上がっており音量が大きくなっている場面であると考えられる．よって今後は，メル周波数ケプストラム係数（MFCC）などを用いた音に関する新たな特徴量を導入し，精度向上を図る．また，「笑顔」のみの精度が低いことに関して，「会話風景として残したい場面」と笑顔の度合いには関連が低い可能性があることが示唆される．

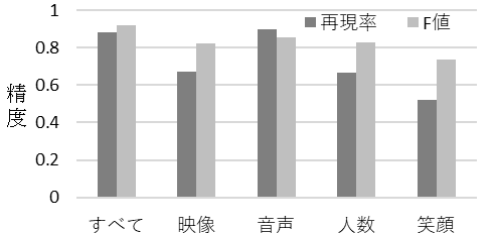

図 3　特徴による分類精度の比較

5. 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のうち，特に「会話風景」について述べ，場面判定に用いる分類器のオフライン評価を行った．「会話風景」の判定を行う分類器の精度には改善の余地があるが，「会話風景」と判定された区間と不正解となった区間には視覚的類似性が認められた．よって，視覚的・時間的な距離を考慮した処理の追加による精度向上が見込まれる．また，笑顔や人数と比較して，音量が重要であることが示唆された．今後の課題として，「会話風景」区間との類似度を考慮した判定の追加や音声に関する新たな特徴量の検討が挙げられる．

謝辞

本研究の一部は科研費補助金ならびに文科省特別経費「持続可能社会に向けた知的情報空間技術の創出」の支援を受けた．

参考文献（web ページは全て 2017-01-09 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, 相澤; “ライフログビデオのためのコンテキスト推定”, 信学技報 IE, 画像工学, 103.514, pp.67-72, 2003
[4] 中村；“映像によるライフログ”, 情報の科学と技術, pp.57-62, 2013
[5] 木下, 小坂, 藤波; “思い出の楽しい振り返りのための身体装着型カメラによる体験自動記録”, 情処研報, Vol.2016-UBI-51, No. 7, 2016.
[6] Google, “Cloud Vision API”, https://cloud.google.com/vision/
[7] The University of Waikato “Weka”, http://www.cs.waikato.ac.nz/ml/weka/
[8] Chawla, N. V., et al. “Synthetic Minority Over-sampling Technique.” Journal of artificial intelligence research, 16, pp.321-357, 2002
[9] Chen, C., et al. “Using random forest to learn imbalanced data.” University of California, Berkeley, pp.1-12, 2004.
[10] He, H., et al. “Learning from imbalanced data.” IEEE Transactions on knowledge and data engineering 21.9, pp.1263-1284, 2009.
[11] Hall, M. A. “Correlation-based Feature Selection for Machine Learning.” PhD thesis, The University of Waikato, April 1999.