論文 / 著書情報
Article / Book Information

論題(和文)	話者認識と顔画像認識を用いた映像におけるマルチモーダル人物同定
Title(English)	
著者(和文)	西 史人, 井上 中順, 岩野 公司, 篠田 浩一
Authors(English)	Fumito Nishi, Nakamasa Inoue, Koji Iwano, Koichi Shinoda
出典(和文)	日本音響学会2017年春季研究発表会講演論文集, , , pp. 129-130
Citation(English)	Proc.of ASJ 2017 Spring Meeting, , , pp. 129-130
発行日 / Pub. date	2017, 3

話者認識と顔画像認識を用いた映像におけるマルチモーダル人物同定*

☆西 史人 (東工大)， △井上 中順 (東工大)， 岩野 公司 (都市大)， 篠田 浩一 (東工大)

1 はじめに

顔画像と音声の情報を用いて映像中の人物を同定するマルチモーダル人物同定の一手法を提案する．本研究では，話者認識，顔画像認識，背景の類似度を用いる人物同定を行う．従来手法として，話者認識と顔画像認識を用いた手法 [1]，話者認識と背景類似度を用いた手法 [2] があるが，それらすべてを用いた方法はなかった．

2 提案手法

人物同定とは，映像の各ショット (カメラの切り替わりから次の切り替わりまでの区間) における人物の名前を同定することである．ここで，人物の発話がない場合は同定をしない．ショット内で複数の人物が現れる場合はそれら全てを同定する．人物名の情報は画面上の文字に対する文字認識 (OCR) により自動獲得する．

2.1 話者認識

映像全体に対し，Rouvier ら [3] の手法で音声検出を行い，得られた発話区間から i-vector [4] を算出する．発話がショットをまたがっている場合があるが，i-vector はショット単位ではなく，発話単位で算出する．

2.2 顔画像認識

映像全体に対して Danelljan ら [5] の手法で顔検出と顔追跡を行い，顔区間を求める．検出された顔区間の各フレームの静止顔画像を FaceNet (畳み込みニューラルネットワーク) [6] に入力して出力を得る．そして顔区間のすべてのフレームの出力を用いて i-vector を計算する．2.1 と同様，i-vector はショットの境界によらず，顔区間単位で算出したものを用いる．

2.3 背景認識

各ショットにおける画像全体の HSV から，色相のヒストグラム (16bin) を求める．

2.4 文字認識 (OCR)

映像全体に対し，OCR を行う．ただし，OCR 結果は Le ら [7] による．一つのショットに複数の名前が検出されることや，同じ名前が別のショットで検出されることがある．

2.5 類似度

話者スコア A，顔画像スコア F は，i-vector 間のコサイン類似度を 0 と 1 の間の値に正規化したものを用いる．また，ショット間の背景スコア B は，色相ヒストグラムの相関を正規化した値である．

2.6 対応付け

まず，OCR 結果が存在するショットに対して処理を行う．一般に一つのショットに複数の人物名，発話区間，顔区間がある．各々の発話区間や顔区間に対して，複数の人物名全てを対応付ける．同じ名前の OCR 結果をショットが複数ある場合，その全ての人物名と各発話区間，各顔区間を対応付ける．

次に，OCR 結果のないショットの発話区間に，話者スコアが最大となる人物名を対応付ける．顔区間についても同等の処理で対応付けを行う．

2.7 信頼度

各ショットにおける各人物の信頼度 C を，話者スコア，顔画像スコア，背景スコアを用いて Late Fusion 法で計算する．信頼度は次の式で表される．

$$C = (1-\alpha)A + \alpha(1-\sigma)F + \alpha\sigma\beta B \qquad (1)$$

α は話者スコアと顔画像スコア，背景スコアの重み，β は顔画像と背景画像の信頼度を調整するための重み，σ はショット s で顔が検出されず，発話が検出された場合 1，それ以外の場合に 0 となる変数である．一つのショット内で同じ人物の発話区間，顔区間が複数存在する場合，話者スコア，顔画像スコアがそれぞれ最大となる区間のスコアを用いる．

---

*Multimodal person identification using speaker recognition and face recognition. by Fumito Nishi (Tokyo Institute of Technology), Nakamasa Inoue (Tokyo Institute of Technology), Koji Iwano (Tokyo City University) and Koichi Shinoda (Tokyo Institute of Technology)

融合法	MAP@1	MAP@10	MAP@100
EUMSSI [7]	0.792	0.652	0.631
A	0.434	0.331	0.320
F	0.629	0.460	0.451
B	0.552	0.424	0.413
A + F	0.699	0.564	0.540
A + B	0.562	0.434	0.421
F + B	0.672	0.503	0.492
A + F + B	0.709	0.578	0.563

Table 1　各手法を融合した際の Mean Average Precision @m (MAP@m)

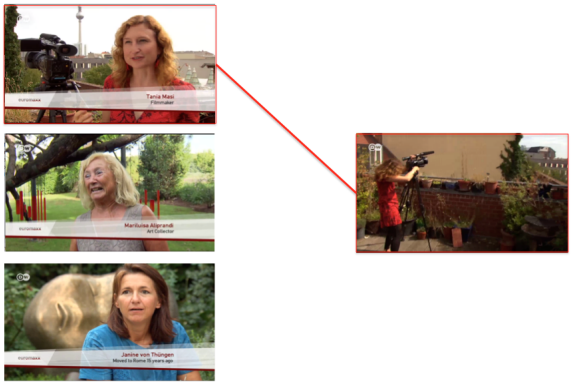

Fig. 1　背景による改善例:人物名が検出されたショット (左) と検出されなかったショット (右)

3　実験

3.1　実験条件

実験データセットには，国際ワークショップである MediaEval2016 における人物同定を行うタスク，Person Discovery in TV の映像データを用いる．データセットのテストデータはニュースやドキュメンタリー，トークショーなどからなる合計 153 時間の映像であり，開発データはそれとは異なるニュース番組からなる 106 時間の映像である．話者，顔画像の Universal Background Model (UBM) と全変動行列は，開発データを用いて学習した．評価基準としては各人物を信頼度順に m 個検索し，全体としての精度を表す Mean Average Precision @m (MAP@m) を用いる．

比較として Le ら [7] を用いた．これは，話者特徴量として i-vector，顔特徴量として，顔画像を複数領域に分割し，各領域で求められた DCT を i-vector 化したベクトルを用いた手法である．

3.2　実験結果

表 1 は各スコアを融合した際の MAP を示している．結果から，話者スコアに加え顔画像スコアを用いることにより精度が高くなることがわかる．また，話者スコアと顔画像スコアに背景スコアを追加した場合においても改善が見られる．これは，図 1(右) のように顔検出が難しいショットにおいても，図 1(左) のように，人物名が現れている背景と比較することで人物同定ができたためである．また，Le ら [7] と比較した場合，MAP@1 の結果は 8.3 ポイント下回っている．これは，Le らは異なる複数の OCR 手法を用いて，より良い Recall を実現しているためである．

4　まとめ

本研究では話者，顔，背景情報を利用したマルチモーダル人物同定手法を提案した．今後の課題として，複数の OCR 結果を併用することによる性能改善が挙げられる．

参考文献

[1] Nam Le et al. Eumssi team at the mediaeval person discovery challenge. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, No. EPFL-CONF-213706, 2015.

[2] Meriem Bendris et al. Percolatte: A multimodal person discovery system in tv broadcast for the mediaeval 2015 evaluation campaign. 2015.

[3] Mickael Rouvier et al. An open-source state-of-the-art toolbox for broadcast news diarization. Technical report, Idiap, 2013.

[4] 小川哲司, 塩田さやか. i-vector を用いた話者認識. 日本音響学会誌 70 巻 6 号, pp. 332–339, 2014.

[5] Martin Danelljan et al. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[6] Florian Schroff et al. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on CVPR*, pp. 815–823, 2015.

[7] Nam Le et al. Eumssi team at the mediaeval person discovery challenge 2016. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*, No. EPFL-CONF-223040, 2016.