固定カメラ映像からの音声・画像情報を用いた映像コンテンツの生成

足立　順†　　滝口　哲也††　　有木　康雄††

† 神戸大学大学院自然科学研究科　〒 657-8501 神戸市灘区六甲台町 1-1
†† 神戸大学自然科学系先端融合研究環　〒 657-8501 神戸市灘区六甲台町 1-1
E-mail: †j-adachi@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

あらまし　固定ビデオカメラで撮影された映像から，音声情報と画像情報を用いて，ダイジェスト映像を生成する研究を行っている．本研究では，一般家庭でのホームビデオ撮影や，車内映像，会議映像といった閉鎖された空間での撮影を想定し，撮影者に技術的・肉体的負担を与えない固定ビデオカメラ映像から，興味深く面白い映像を生成することを目的とする．本稿では，音声情報，画像情報を用い，会話シーンを中心とした映像編集を提案する．会話シーンの抽出は，低 SNR 環境下においても頑健な音声/非音声の区間検出が可能な AdaBoost に基づく手法で音声・非音声を判別し，各会話シーンセグメントを蓄積していく．蓄積された各会話シーンセグメントごとに，2ch(ステレオ) マイク間の信号到来時間を CSP 法 (Cross-Power Spectrum Phase Analysis) を用いて推定し，その値より発話方向を推定する．発話方向を推定した後，ディジタルカメラワークにより発話者へのズームを行う．この際，発話者の顔を中心にズームを行うために，OpenCV (Intel Open Source Computer Vision Library) による顔画像検索を行い，それによって得た座標の値に応じてズームイン・ズームアウトのカメラワークを決定する．以上のようなシステムを提案・実験を行い，ダイジェスト映像を生成した．

キーワード　固定カメラ映像，発話検出，発話方向，ディジタルカメラワーク，顔画像検索，コンテンツ生成

Image Content Generation Using Voice and Image Information from Fixed Camera

Jun ADACHI†, Tetsuya TAKIGUCHI††, and Yasuo ARIKI††

† Graduate School of Science and Technology, Kobe University　Rokkodai-cho 1–1, Nada-ku, Kobe-shi, Hyogo 657–8051 Japan
†† Organization of Advanced Science and Technology, Kobe University　Rokkodai-cho 1–1, Nada-ku, Kobe-shi, Hyogo 657–8051 Japan
E-mail: †j-adachi@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

Abstract　In digital videos audio has a key index, that can provide useful information for home video retrieval, such as capturing conversations only, clipping only talking people. In this paper, we propose home video editing system based on audio with two-channel (stereo) microphone that is standard equipment on video cameras, where the video content is automatically recorded without a cameraman. In order to capture only a talking person on video, a novel voice/non-voice detection algorithm using AdaBoost, which can achieve extremely high detection rates in noisy environments. In addition, the sound source direction is estimated by the CSP (Crosspower-Spectrum Phase) method in order to zoom in/out talking person by clipping frames from videos, where a two-channel (stereo) microphone is used to obtain information about time differences between the microphones.

Key words　Video editing, audio, voice detection, sound source direction, digital camera work,

1. はじめに

近年，コンピュータの小型化，記憶デバイスの大容量化により個人の行動記録の入手が簡単になっている．Microsoft 社が推進している MyLifeBits Project [1] に代表されるような，日常生活の情報を記憶・蓄積していく試み（ライフログ）が盛んに行われている．一般家庭でも，小型化され安価になったディジタルカメラ・ディジタルビデオカメラを用いて，日常生活や

パーティ等の映像を記録する事が多く行われている．だが，このような一般家庭で撮影する場合，撮影者に時間的・体力的な浪費を負わす上に，撮影を行っている人物が画面上に登場しないと言う問題を抱えている．そこで，撮影者に負担を与えない固定カメラによる撮影が行われるが，そのように撮影された映像は，ズームといったカメラワークが無いため，どうしても単調な映像になってしまう．さらに，冗長で不必要なシーンが多くなり，視聴の際に編集や検索などの手間をかけさせてしまう事になる．このような背景から，それらの冗長な映像から映像編集を行う研究が盛んになっている [2] [3] [4] [5].

これまでの多くの研究では画像情報を用いた編集がなされてきた．しかし，そのような画像情報を基にした編集では，動きの少ない会話シーンや，横顔や下を向いた顔など顔認識が行えないシーンなどで，重要なシーンが欠落してしまう可能性がある．

そこで，本研究では音声情報と画像情報を組み合わせて，会話シーンを中心とした映像編集を提案する．まず，2チャンネルマイクロフォンで得た音声情報を元に，得られた映像から人物の発話区間を抜き出し，会話シーンを抽出する．そして得られた会話シーンセグメントごとに，発話方向を推定し，ズームイン・ズームアウト等のカメラワークを決定する．ズームによって切り出される範囲は，顔検出情報を用いて得られた発話者の顔を中心に切り出す．撮影・編集といった負担無く，興味深い映像コンテンツを作成し提示することを目標とする．

2. 提案手法の概要

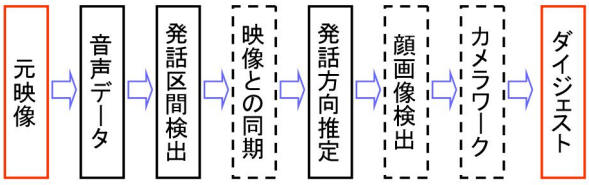

図 1　提案手法における処理の流れ

本システムにおける処理の流れを図 1 に示す．得られた音声データから発話区間を検出し，発話区間のみの映像を抽出する．次に発話方向推定を行い，発話方向に応じたズームイン・ズームアウト等のカメラワークを行う．以下，3 章において発話区間検出，4 章において発話方向推定，5 章においてカメラワークについて述べる．

3. 発話区間検出

発話区間検出については，これまでに我々が提案している低 SNR 環境下においても頑健な音声/非音声の区間検出が可能な AdaBoost に基づく手法 [6] を用いた．

3.1 Real AdaBoost による音声/非音声の識別

AdaBoost は単純な識別器を複数組み合わせることによって，精度の高い識別器を構成する Boosting 法の中でも顕著な性能を示す手法である．今回は AdaBoost をさらに発展させた，Real AdaBoost を用いて音声/非音声の判別を行った．以下にアルゴリズムの詳細を述べる．

学習データ数を n，繰り返し回数を M とする．これらの値はあらかじめ決定しておく必要がある．学習データを $x_i(i = 1, \cdots, n)$，各データの重みを $w_i(i = 1, \cdots, n)$ とする．x_i としては音声の各フレームから取り出された MFCC(Mel Frequency Cepstrum Coefficient) を用いる．各データ x_i には，あらかじめ $y \in -1, +1$ を与えておく．すなわち，データ x_i が音声であれば $y = +1$，データ x_i が非音声であれば $y = -1$ とする．

（1） 各データの重みを $w_{1i} = \frac{1}{n}$ で初期化する．

（2） $m = 1, \cdots, M$ で以下を実行する．

(a) w_{mi} を確率分布として，x_i から重複を許して n 個，重み付きリサンプリングしたものを $x_i^{'}$ とする．

(b) $x_i^{'}$ に対して弱識別器 $f_m(x)$ を構成する．弱識別器 $f_m(x)$ は信頼度を基に生成するものでなければならない．本研究では弱識別器として，CART による 2 分木を用いた [7].

(c) こうして得られた弱識別器 $f_m(x)$ を用いて $c_m(x_i)$ を得る．

$$c_m(x_i) = \frac{1}{2}\log\left(\frac{f_m(x_i)}{1 - f_m(x_i)}\right) \tag{1}$$

(d) 各データ x_i の重みを $w_{(m+1)i}$ に更新する．

$$w_{(m+1)i} = \frac{w_{mi}e^{-y_i c_m(x_i)}}{\sum_{r=1}^{n} w_{mr}e^{-y_r c_m(x_r)}} \tag{2}$$

（3） 最終的な出力として強識別器 $F(x)$ を得る．

$$F(x) = \sum_{m=1}^{M} c_m(x) \tag{3}$$

通常，$sign$ により出力を $-1, +1$ にするが，ここでは各フレームごとの信頼度を用いるため，$\sum_{m=1}^{M} c_m(x)$ により算出された値をそのまま用いることにした．閾値 α を決定し，入力データ x に対し $F(x) \geqq \alpha$ であれば音声，$F(x) < \alpha$ であれば非音声とする．AdaBoost の重要な性質として，誤ったデータに対するサンプリング重みを増し，次回以降の学習でそれらのデータを重点的に学習する，と言うことが挙げられる．それにより，前段の弱識別器が誤識別を起してしまったデータに対しても，後段の弱識別器が正しく識別するため，最終的に正しい識別結果を得ることができる．

3.2 音声区間検出

音声区間が分断されることを避けるため，式 (4) により，隣接する n フレーム間でスムージングを行う．

$$F'(x_i) = \frac{1}{n}\sum_{j=i-\frac{n}{2}}^{i+\frac{n}{2}} F(x_j) \tag{4}$$

得られた $F'(x_i)$ が，閾値 α 以上であるならば音声，α 以下であるならば非音声とし，暫定的な音声区間を得る．こうして得られた音声/非音声の区間から，連続時間が短いものを取り除くことにより，最終的な音声区間を得る．

こうして得られた各音声区間を，元映像と同期を取って会話シーンセグメントとして切り出す．切り出された各セグメントは，図 2 のように蓄積していく．

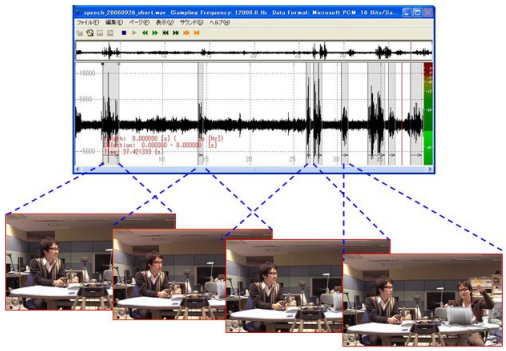

図 2　会話シーンセグメントの蓄積

4.　発話方向推定

発話方向の推定については，2 チャンネルマイクロフォンの到来時間差を，相互相関の一種である CSP(Cross-Power Spectrum Phase) 係数 [8] に基づいて獲得し，その値から発話方向を推定する [9] [10].

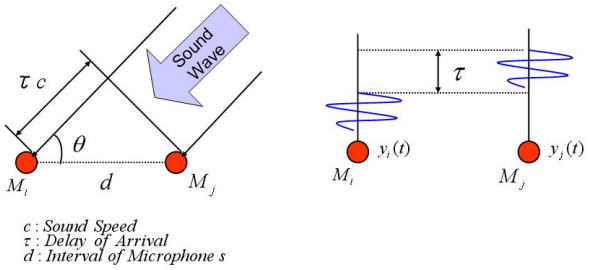

図 3　2 チャンネルマイクロフォン間の信号到来時間差

図 3 に 2 チャンネルマイクロフォンにおける，信号到来時間差の概要図を示す．2 つのマイクロフォン M_i，M_j で受音した信号をそれぞれ $y_i(t)$，$y_j(t)$ とすると，CSP 係数 $CSP_{i,j}(k)$ は，

$$CSP_{i,j}(k) = IDFT[\frac{DFT[y_i(t)]DFT[y_i(t)]}{|DFT[y_j(t)]||DFT[y_j(t)]|}] \qquad (5)$$

で表される．

到来遅延時間 τ は，

$$\tau = arg\max_{k}(CSP_{i,j}(k)) \qquad (6)$$

として推定される．また，発話方向 θ は次式 (7) により得られる．

$$\theta = \cos^{-1}(\frac{c \cdot \tau / f}{d}) \qquad (7)$$

(ただし c は音速，f は標本化周波数，d はマイク間距離．)

図 4〜図 6 はそれぞれ人物 A(60 °付近に存在) の発話セグメント，人物 B(105 °付近に存在) の発話セグメント，A と B の 2 人が発話しているセグメントでの CSP 係数 $CSP_{i,j}(k)$ の各値を示している．(横軸は発話方向 θ)

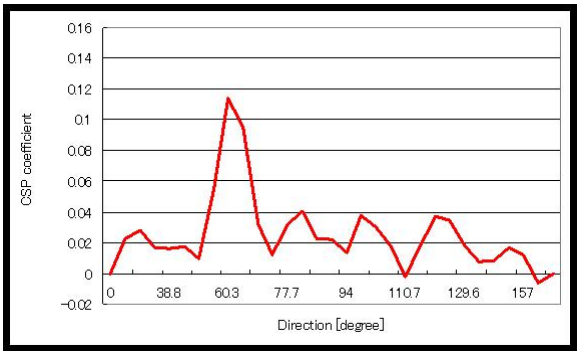

図 4　人物 A の発話セグメントでの CSP 係数

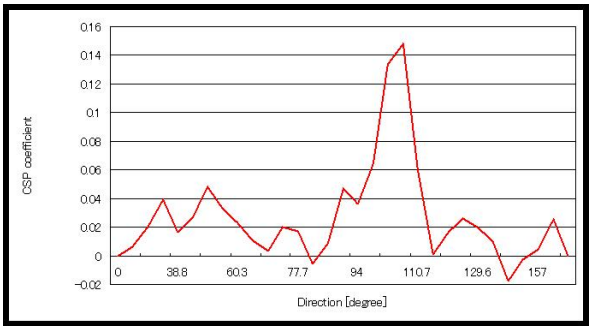

図 5　人物 B の発話セグメントでの CSP 係数

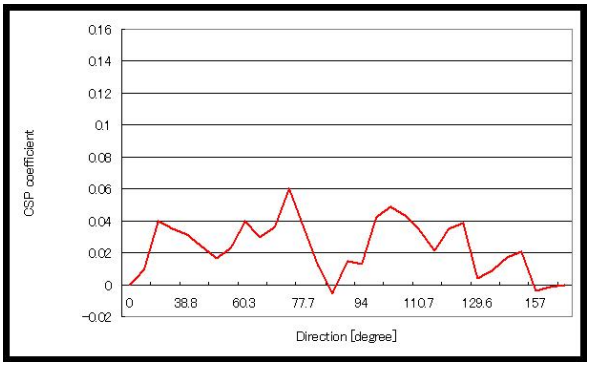

図 6　2 者会話セグメントでの CSP 係数

5.　カメラワーク決定

発話方向の推定結果を基に，ズームイン・ズームアウトのカメラワークを決定する．カメラワークにはディジタルカメラワーク技術を用いる．ディジタルカメラワークとは，高解像度の固定カメラにより映像を撮影し，各フレームをディジタル処理してクリッピングを行い，擬似的なカメラワークを実現するものである．ディジタルカメラワークの利点は，最初に撮影した高解像度映像があれば何度でも映像を編集し直せることである [11]．特定の嗜好にあわせた映像に編集を変更することが可能であり，一つの元映像から，各個人の嗜好に特化した映像が生成できる．

図 7 にカメラワーク決定の処理の流れを示す．まず，発話区間検出によって得られた各会話シーンセグメントごとに発話方向を推定する．一つの発話セグメント内で，複数人 (複数の角度) が発話している場合，式 (5) で得られる CSP 係数 $CSP_{i,j}(k)$

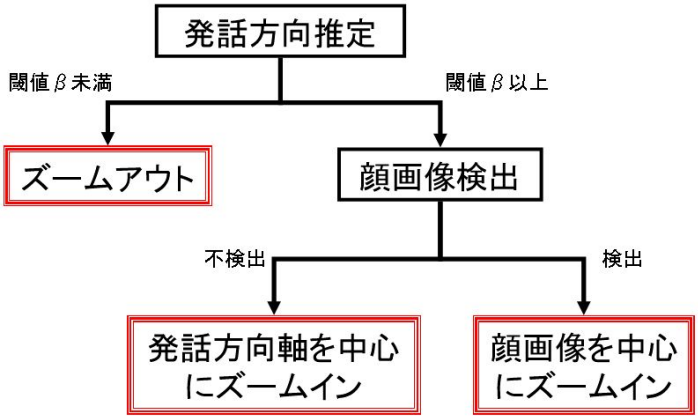

図 7　カメラワーク処理の流れ

は低くなるため，適当な閾値 β を設け，$CSP_{i,j}(k) \geqq \beta$ ならば発話者が一人としてズームイン処理，$CSP_{i,j}(k) < \beta$ ならば発話者が複数として全体を写すカメラワークに決定する．

会話シーン内において最も興味をもたれるのは，発話人物の顔である．そこで顔検出を行い，発話方向付近にいる人物の顔座標を中心にズームイン映像を生成する．また，発話人物が横や下を向いていて，発話人物の顔が検出できない場合は，発話方向軸を中心に映像をクリッピングする．本研究では，OpenCV (Intel Open Source Computer Vision Library) を用いて顔検出を行った [12].

6.　実験と評価

以下の図 8 で示す環境で撮影した映像を実験に用いた．

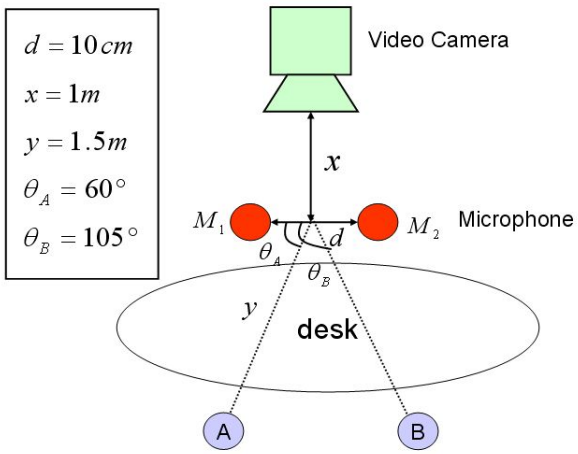

図 8　実 験 環 境

被験者は男性 2 名 (被験者 A：60° 付近に存在，被験者 B：105° 付近に存在) で，イスに座っての会話シーン，時間は 303 秒撮影した．撮影に用いたカメラは，解像度の高いハイビジョンカメラ (Victor GR-HD-1) を用いた．解像度は 1280× 720(規格名:720p) のとき，カメラの画角 ω[deg] は，

$$\tan\frac{\theta}{2} = \frac{X}{2f}, X = (h, w, d) \tag{8}$$

$$\omega = \frac{180}{\pi} \times 2\tan^{-1}(\frac{X}{2f}) \tag{9}$$

で表される．ただし，X は撮像エリア寸法であり，h は垂直方向 (2.735 mm)，w は水平方向 (4.864 mm)，d は対角方向 (5.580 mm) でそれぞれ表され，また f はレンズの焦点距離 (5.2 mm) である．

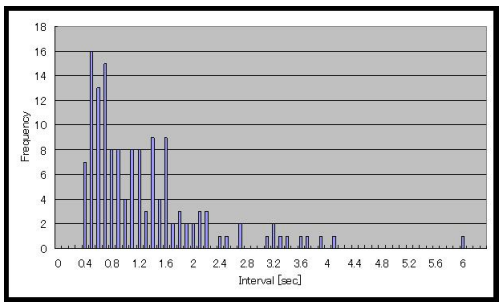

図 9　会話シーンセグメント時間長

図 9 に，AdaBoost を用いた発話区間検出によって切り出された会話シーンの，セグメント時間長グラフを示す．切り出されたセグメントは 141 箇所 (186.4 秒) であり，116.6 秒間 (61.4%) 映像が短縮された．セグメントの最長は 6.07 秒，最短は 0.46 秒であり，検出精度は 94.6% であった．誤検出の原因としては，雑音環境の学習不足が挙げられる．また，セグメントの時間長が短すぎて，発話かそうでないか判別しがたいものがあったため，それらは誤検出とした．セグメント時間長を，一定の時間長以上に制限することによって，改善が得られると考えられる．

発話者を特定し，カメラワークを決定するために，シーンごとに切り出され蓄積された各会話シーンセグメントに CSP を用いて，発話方向を推定する．発話区間検出によって切り出された各セグメントを，人間の耳で評価し正解タグを付与する．正解タグとしては，

（1） A 単独発話シーン (セグメント内で A のみが発話)
（2） B 単独発話シーン (セグメント内で B のみが発話)
（3） 2 者会話シーン (セグメント内で A，B の両者が発話)

の 3 つを用意する．閾値 β を変えて，正答率の評価を行った．表 1 では検出された区間単位での正答率を，表 2 では検出された時間単位での正答率をそれぞれ示している．

表 1　区間正答率

閾値 β	1(ズーム無し)	0.1	0.08	0(常にズーム)
全区間数	141	141	141	141
正解区関数	50	101	103	91
正答率	35.5%	71.6%	73.1%	64.5%

表 2　時間正答率

閾値 β	1(ズーム無し)	0.1	0.08	0(常にズーム)
全時間長	186.5	186.5	186.5	186.5
正解時間長	84.6	133.9	120.3	101.9
正答率	45.4%	71.8%	65.5%	54.6%

ただし，表において，閾値 β が 1 の場合は常にズーム無しの映像，閾値 β が 0 の場合は常に A か B へのズーム映像である．

ズーム無しの映像に比べて，正答率が上がっており，発話者に注目した編集が行えることが示された．判別失敗事例はタグ (1) とタグ (3) の判別誤りか，タグ (2) とタグ (3) の判別誤りであり，タグ (1) とタグ (2) での判別誤りは見られなかった．その理由としては，同一セグメント内での 2 者の発話バランスなどが考えられるが，最適な閾値の設定を今後の課題としたい．以下の表 3 において，2 者会話シーンでの同時発話 (2 者の発話が重なっている) と交互発話 (2 者の発話が交互になっている) の割合を，表 4 において，それぞれの正答率を示している．

表 3　2 者会話シーンにおける内訳

	同時発話	交互発話	総区間
区間数	17	30	47
割合	36.2%	63.8%	100%

表 4　2 者会話シーンにおける正答率

閾値 β	0.1	0.08
同時発話正答率	88.2%	70.6%
交互発話正答率	76.7%	56.7%
2 者発話正答率	80.9%	61.7%

図 10 で発話区間検出の例，図 11 で発話方向判別の例を示す．

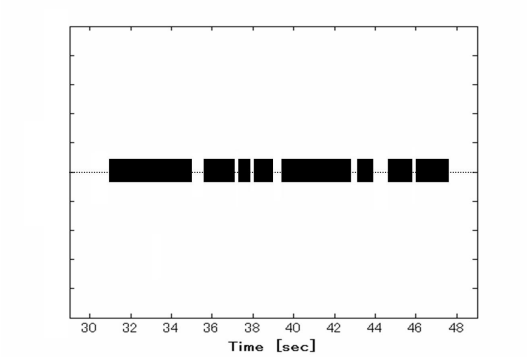

図 10　発話区間検出の例

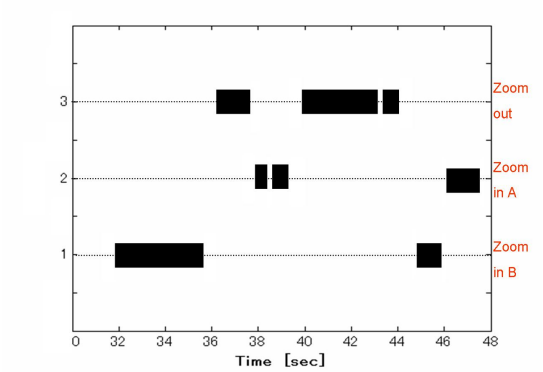

図 11　発話方向判別の例

次に，OpenCV による顔検出を行う．発話方向推定で発話者 A，または B へのズームに判定された各セグメントにおいて，CSP 係数の最も高かった角度付近の顔検出を行う．実験の映像は 29 フレーム/秒で行い，各セグメントごとで顔中心座標の平均座標を求め，その点を中心に映像を切り出し拡大した．また，横や下を向いており発話者の顔が一定以上認識できない場合は，発話軸と y 軸方向の中心線の交点を中心点として切り出し，ズーム映像を生成した．今回の実験では各セグメントごとにカメラワークを決定して，セグメント内でのカメラワークは行っていない．理由は短時間でズームやパン (平行移動) により，映像が見難くなってしまうのを避けるためである．

以下において，作成された映像の例を示す．図 12 では発話者 A へのズーム映像，図 13 では発話者 B へのズーム映像，図 14 ではズーム無し映像をそれぞれ表している．

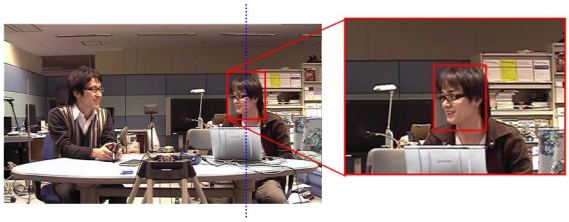

図 12　カメラワークの例:発話者 A へのズームイン

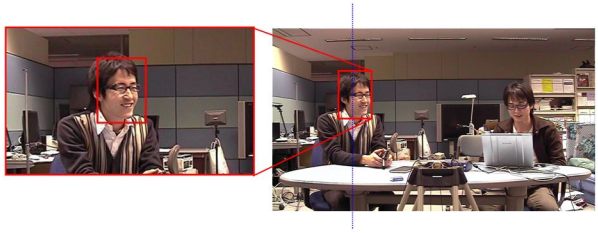

図 13　カメラワークの例:発話者 B へのズームイン

図 14　カメラワークの例:ズームアウト

7. ま と め

本論文では，固定カメラによって撮影された長時間の映像から興味深い映像コンテンツを生成する方法として，音声情報と画像情報を用いて，ダイジェスト映像を生成する手法を提案した．2 チャンネルマイクロフォンから得られた音声情報を用いて発話区間のみを切り出し，切り出した各会話シーンセグメントごとに発話方向と顔画像検出を用いることによってディジタルカメラワークを決定する方法を提案し，実験を行った．その結果，冗長な映像からのダイジェスト映像を生成することができた．

今後の課題としては，まず発話区間検出の精度向上が挙げられる．AdaBoost を用いた検出では事前に音声・非音声の学習が必要であり，また過学習をしてしまうと実験環境に依存してしまうため，汎用性を高めるためには，この問題を解決したい．

また，生成された映像については以下のような課題が挙げられる．まず，元の映像と比較した評価結果を求めることである．現在は生成した映像について主観で評価を行っているため，客観的に評価を行うことが必要である [13]．また，主観的な評価ではあるが，ズームイン・ズームアウトの切り替えが一瞬で行われるために見難さが生じる場合があり，さらにシーン間の結合部分が不自然であると感じた．そのようなことを含めて，プロのカメラマンの撮影した映像を参考に，パン (カメラの平行移動) 等を含めた視聴者にとって見やすく飽きないためのカメラワークの向上を目指したい．

文　献

[1] http://research.microsoft.com/barc/MediaPresence/MyLifeBits.aspx

[2] HUA, X. S., et al. AVE: automated home video editing, Proceedings of the eleventh ACM international conference on Multimedia , pp. 490 - 497 ,2003.

[3] HUA, X. S., et al. Automatic music video generation based on temporal pattern analysis, Proceedings of the 12th annual ACM international conference on Multimedia , pp. 472 - 475 ,2004.

[4] P. Wu. A semi-automatic approach to detect highlights for home video annotation, Proc. ICASSP, pp. 957 - 960 ,2004.

[5] B. Adams, S. Venkatesh. Dynamic shot suggestion filtering for home video based on user performance, Proceedings of the 13th annual ACM international conference on Multimedia , pp. 363 - 366 ,2005.

[6] 松田博義，滝口哲也，有木康雄，“Real Adaboost による音声区間検出，” 日本音響学会 2006 年秋季研究発表会，2-P-12, pp.117 － 118，Sep 2006.

[7] http://research.graphicon.ru/generalprojects/about-us.html.

[8] M. Omologo, P. Svaizer. Acoustic source location in noisy and reverberant environment using CSP analysis, Proc. ICASSP, pp. 9217 - 924 ,1996.

[9] 横江優貴, 伊藤義道, 馬場口登，“マルチメディアログ作成のための音声による話者方向推定，” 画像の認識・理解シンポジウム (MIRU2006)，pp.1145 － 1151，Jul 2006.

[10] F. Asano, J. Ogata. Detection and Separation of Speech Events in Meeting Recordings, Proc. Interspeech 2006, pp.2586-2589, Sep. 2006

[11] 窪田進太郎，有木康雄，熊野雅仁，“ディジタルカメラワークを用いたボールと選手の状況認識に基づくサッカー映像の自動生成，” 画像の認識・理解シンポジウム (MIRU2005)，pp.1145 － 1151，Jul 2005.

[12] http://www.intel.com/technology/computing/opencv/index.htm.

[13] 片山貞昭，“AHP によるシステム評価，” 情処研報告，IPSJ, Oct 1990.