4J−4

協調型ドラマシーン理解システムによるシーン、カット、音声の対応付け（実験）

影山誠　柳沼良知　坂内正夫
東京大学生産技術研究所

1　はじめに

近年、マルチメディアに関する研究が盛んに行なわれている。メディアの認識理解技術は、そのメディアの変換, 加工, 編集をおこなう際に重要な役割を果たすことが予想されるが、メディアを単独に認識, 理解しようとする場合、十分な理解が行なえないことも多い。この点を改良する方法として、複数のメディアからの情報を融合することによって、より高度な認識を行なうことが考えられる。我々は、このような試みの1つとして、認識の対象をドラマシーンに絞り、映像、音声、シナリオなどから得た情報を統合して認識を行なうシステムの開発を行なっている [1]。本システムは、映像、音声、シナリオの3つのメディアを協調させる際に、メディアに依存しない「共通概念」を用いる点に特徴がある。

本稿では、このシステムの動作例として、映像, 音声, シナリオの協調によるシーンとカットの対応付けの実験について述べる。

2　シーンとカットとの対応付け

今回行なうのは、シーンとカットとの対応付けの実験である。ドラマはいくつかのシーンから構成されている。また、シーンはいくつかのカットから構成されている。シーンとカットとの対応付けとは、どのカットがどのシーンに属するかということを認識することである。

シーンはシナリオのシーン番号から抽出できる。

カットの抽出は、画像を RGB から HVC へ MTM アルゴリズムを用いて変換し、色相のヒストグラム (h_x) から
$\chi^2 = \sum_{i=0}^{63}\{h_2(i) - h_1(i)\}^2/h_1(i)$ を求め、$\chi^2 \geq 7000$ のときにカットが変わるとして抽出する。

シーンはシナリオ、カットは画像というようにメディアが異なるので、メディア間での協調、認識を行なう際には何らかの拠り所となるものが必要である。これが「共通概念」である。つまり、それぞれのメディアから共通概念を抽出してそれをもとに認識を行なうことになる。

3　共通概念を用いた対応付け

ここでは共通概念について説明する。今回用いる共通概念は、時間、人数、ABAB、女性の存在、の4つである。

3.1　時間

映像、音声ははもともと実時間のメディアであるから、はじめから時間情報を持っているし、互いに同期している。従って、カットの開始、終了時間を求めるのは容易である。

それに対し、シナリオは時間的な情報を持っていない。従って、その時間を大まかにでも推定する必要がある。そこで、文字数により時間を比例配分し、その時間±２０秒を推定した時間とする。

3.2　人数

これは、そのメディアのなかに登場する人間の数である。
映像では先ほどの HVC において、$270^\circ \leq H \leq 350^\circ$ で４００画素以上の領域をを肌色領域、また、$0 \leq V \leq 4$ で４００画素以上の領域をを黒色領域とし、肌色領域の外接長方形の上辺と黒色領域の外接長方形が重なったものを人としてその数を求める。

シナリオではそのシーンに登場する話者の数を返すようにする。

3.3　ABAB

ABAB とは、例えば会話などで交互に同じ人間が現れるような場面のことを指す。
映像では1番目と3番目のカット, 2番目と4番目のカットが同じ場面のものを選ぶようにする。具体的には、1と3、2と4の画像が$\chi^2 \leq 2000$ の関係を満たすものを選ぶ。

シナリオでは1番目と3番目、2番目と4番目の話者が同じものを選び出す。

Matching between scene, cuts, and sound using multimedia cooperative drama scene recogniton system (experiment)
Makoto Kageyama , Yoshitomo Yaginuma , Maksao Sakauchi
Institute of Industrial Science Universuty of Tokyo

3.4 女性の存在

シナリオでは女性の名前が存在するかどうかを調べる。

画像から女性を抽出するのは難しいので、もともと画像と同期している音声を用いる。カットの始めから、数秒音声を取り出して、その中からボリュームの大きい部分を取り出す。(ドラマなので、声が大きく聞こえるようになっていると予想される) FFT を用いてケプストラムをもとめ、ピッチを抽出し、基本周波数が２３０ Hz 以上４００ Hz 以下であるものを女性であるとした。正確さを増すために、基本周波数の下限はいくぶん高めにしている。ここで行なっているのは、明らかに女性が存在するカットを見つけることであるので、女性が存在するすべてのシーンを抽出できるとは限らない。

4 協調による認識結果

今回実験に用いたのは、９のシーンを含む７分２０秒の映像から２秒ごとに取り込んだ２２０枚の画像で、これから２２のカットを抽出した。なにも制限をつけずにシーンとカットとの対応付けをとろうとすると ${}_{21}C_8 = 101745$ 通りの組合せが考えられるが、

1. シナリオからの大まかな推定時間に当てはまらない組合せを削除する。
2. ABAB の検出によりシーンを特定する。
3. 人数の対応づけをとり、映像中の人数＞シナリオ中の人数となる組合せを除去する

といった、協調処理を行なうことによって、９０通りにまで組合せを減らすことができた(図１)。

図中の１はシーンとカットとの対応づけが確定しているという意味である。また、

4. 女性の存在の有無により、誤った組合せを除去する

ことによって、３６通りにまで組合せを減らすことができた（図２)。

5 むすび

今回の発表では、ドラマシーン理解システムの具体的動作として、共通概念によるメディア協調を用いたカットとシーンの対応付けの実験について述べた。今後は、より高度なドラマシーンの認識理解についての研究を進めていく予定である。

参考文献

[1] 影山, 柳沼, 坂内, “マルチメディア協調型ドラマシーン理解システムによるカットとシーンの対応付け”, １９９２年電子情報通信学会秋期大会

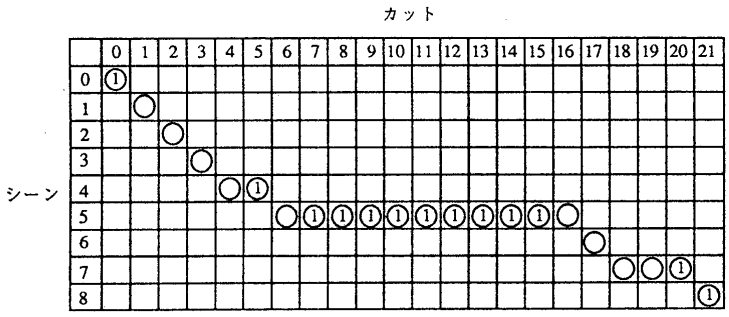

図１　時間、人数、ＡＢＡＢを用いたシーンとカットの対応付け

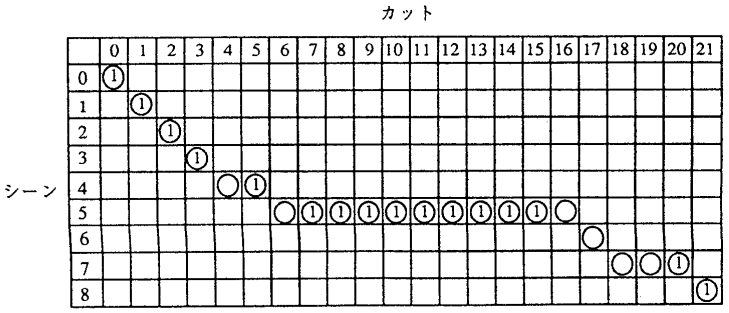

図２　さらに音声を用いたシーンとカットの対応付け