論文 / 著書情報
Article / Book Information

論題(和文)	ホームビデオからのハイライト検出支援のための音声情報の視覚化
Title(English)	
著者(和文)	高木 幸一, 川田 亮一, 篠崎 隆宏, 古井 貞熙
Authors(English)	Ryoichi Kawata, Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会2010年秋季講演論文集, , No. 2-9-11, pp. 69-70
Citation(English)	, , No. 2-9-11, pp. 69-70
発行日 / Pub. date	2010, 9

ホームビデオからのハイライト検出支援のための音声情報の視覚化*

○高木幸一，川田亮一(KDDI 研)，篠崎隆宏，古井貞熙(東工大)

1　はじめに

携帯端末等で撮影されたビデオ（ホームビデオ）を短尺化（要約）する際に，音を聴かず，画面を見るだけで大事な個所（ハイライト）を検出できるようにしたいという要望がある。ところが，音声を聴きながらハイライトを検出する場合とそうでない場合とでは差異が生ずることが予想される。そこで，本稿では，音声情報を聴くことなしに視覚情報(映像＋音声情報を視覚化したもの）のみでハイライト検出を行うことを想定し，その場合に本質的に必要となる，すなわち視覚化すべき音声情報とは何かについて検討したので報告する。

2　音声が映像を補完するケース

本研究では，映像（以下Vと書く）だけからは得られない情報で，かつ音声（以下Aと書く）として本質的な情報は何であるかを見極めることを目標としている。例えばVだけからは得られない情報として，ビデオに映し出されていない物体が発する音の情報はそれに相当する。ただし，すべての音の情報が必要であるとは限らない。そこで，本節では「Vのみ」，「V+A」のそれぞれについて実際に短尺化を行い，その差異を評価する。

2.1 実験

本実験において，被験者は複数のコンテンツに対する短尺化を行う。実験条件はTable 1の通りである。被験者は，

「自分が重要だと考える箇所を含め，その長さ（時間）がオリジナルシーケンスの 20～30%の長さになるように切り出してください（短尺化してください）」

という依頼に対し，N(=5)種類のシーケンスを，「Vのみ」および「V+A」の２通りの情報提示に対して短尺化処理を行う。すなわち，同一シーケンスに対して２度の短尺化処理を行う。あるシーケンスに対して，「Vのみ」の場合を評価する前に「V+A」の場合を評価してしまうと，被験者は「Vのみ」を評価する際に，記憶された「A」の情報をもとに判断してしまう，すなわち，「V+A」との差異が認められなくなる恐れがある。そこで，あえて両者の差異を明確にするために，「Vのみ」の情報を提示した後に「V+A」の情報を提示することとする。また，連続して同一シーケンスを評価することによる飽きを起こさせないようにするため，同一シーケンスは連続して提示しないようにする。以上を踏まえて，シーケンス提示順は「Vのみ」→「V+A」とする。

なお，各被験者に提示するシーケンスはそれぞれ部分的な重なりはあるものの，異なるものとする。

Table 1. Experimental setup

提示シーケンス数／被験者	（全 40 本のシーケンスのうち）５本×２通り (「Vのみ」，「V+A」)
被験者数	16 名（20～40 代・男女）
シーケンス仕様	携帯端末で撮影したもの 時間：30 秒～2 分 File format：MP4（3g2/3gp） Video：H.264，15fps，QCIF～QVGA，64～256kbps Audio：AMR-NB (8kHz, mono, 12.2kbps)
シーケンス内容	一般の方により撮影されたホームビデオ 例：公園での散歩，食事の風景，ペットとの戯れ，ドライブ，ランドマーク旅行など

2.2 結果および考察

実験の結果，「Vのみ」の場合と「V+A」の場合で，当初の予想通り，差異が生じることが確認された。

以上の結果を分析するために，提示したシーケンスに対し，あらかじめオーディオ情報のカテゴリ分けを手動で行っておき，本実験により得られた結果と照合した(Fig. 1)。結果

*Audio information visualization for supporting highlight extraction from home video, by TAKAGI Koichi, KAWADA Ryoichi (KDDI), SHINOZAKI Takahiro and FURUI Sadaoki (Tokyo Institute of Technology).

を Table 2 に示す。本節で「カテゴリ」とは，各セグメント（1秒単位）に対し Table 2 の Sound category に相当する情報が存在するか否かを重複を許して示したものである。「V+A」の結果に対し，「V のみ」の結果で挿入されている箇所(Ins.)，削除されている箇所(Del.)，およびどちらにおいても選択された箇所(Co-corr.)を求め，カテゴリごとに，当該区間に含まれるセグメント（1 秒）数をカウントした。

同表より，「V のみ」の場合と比較して，「V+A」の場合に抽出されたセグメントに含まれている情報として，「人の声」が挙げられることがわかる。特に，「人の声」では Del. が非常に多くなっており，音声を聴くと「人の声」の部分を含めたくなることがわかる。一方，Ins.，Del.両者の差異が大きいものとして「突発音」（machine noise，impulse sound）が挙げられる。突発的な音はハイライトとして意味があるケースもあるが，逆に突発な音が入ることを嫌うケースも考えられる。それゆえ，このような結果になったと想像できる。逆に，オーディオを用いた短尺化の際に，一般的に最も注目されるのは「歓声」の部分である[1]と言われているが，「A」情報が加わっても，「歓声」が抽出される割合が高くなっていない。この理由として，通常の放送用の映像と比較し，ホームビデオでは歓声部分を意図して撮影しようしないこと，また，Co-corr. の数からもわかる通り，歓声があがっている箇所は「V」だけからでも認識できるためであることなどが考えられる。いずれにせよ，今回の目的においては不要であることがわかる。

以上の結果から，「人の声」「突発音」の 2 つの情報が他の情報とは突出して変化があり，これらの情報を効率的に検出し，「V」と一緒に提示することができれば，当初の目的を達成することができると期待できる。

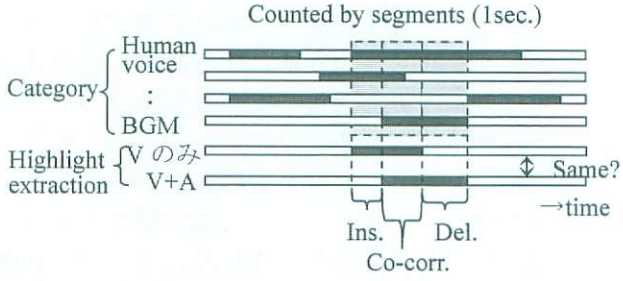

Fig. 1 Evaluation method for highlight extraction results by “only V” and “V+A.”

3 おわりに

ホームビデオを対象とし，音声情報を聴くことなしに視覚情報（映像＋音声情報を視覚化したもの）のみでハイライト検出を行う場合に，本質的に必要となる音声情報とは何かについて検討した。そのため，単一のビデオに対し，音声を聴かない場合と聴いた場合の両者でハイライト検出を行い，そこで選択された箇所の違い（Table 2 の Del.と Ins.）について分析した。その結果，「人の声」と「突発音」の箇所に大きな違いがあり，これらを視覚化して表示する重要性が明らかになった。今後はこれらの音の効率的検出法および視覚化法について明らかにしていく。

参考文献

[1] Otsuka et al., IEEE Trans. on Consumer Electronics 51 (1) 112–116, 2005.

Table 2 Classified number of extracted highlight segments for each sound category.

Sound category	Description	# of segments	Del.	Ins.	Co-corr.
Human voice	人の声	1316	**234(17.8%)**	62(4.7%)	259(19.7%)
Machine noise	機械的な音（ベルなど）	248	**22(8.9%)**	**32(12.9%)**	51(20.6%)
Vehicle	乗物から発せられる音	315	10(3.2%)	12(3.8%)	52(16.5%)
Animal	動物から発せられる音	143	8(5.6%)	2(1.4%)	13(9.1%)
Cheering	歓声	621	21(3.4%)	23(3.7%)	101(16.3%)
Applause	拍手	321	18(5.6%)	12(3.7%)	71(22.1%)
BGM	背景に流れる音楽（除楽器）	1437	65(4.5%)	105(7.3%)	72(5.0%)
Musical instrument	楽器の音全般	312	12(3.8%)	2(0.6%)	23(7.4%)
Water, wave	水に関連する音	102	2(2.0%)	0(0.0%)	9(8.8%)
Wind	風の音	82	0(0.0%)	3(3.7%)	8(9.8%)
Life sound	人のざわつき，足音	1050	13(1.2%)	12(1.1%)	35(3.3%)
Impulse sound	モノがぶつかる音等	612	**39(6.4%)**	**41(6.7%)**	42(6.9%)