論文 / 著書情報
Article / Book Information

論題	多段SVMを用いた頑健な動画ショット境界検出
Title	
著者	宮村 祐一, 中村 太一, 篠田 浩一, 古井貞熙
Author(s)	Yuichi Miyamura, Taichi Nakamura, Koichi Shinoda, SADAOKI FURUI
出典	画像の認識・理解シンポジウム（MIRU 2007）IS-2-19, Vol. , No. , pp. 815-820
Citation	, Vol. , No. , pp. 815-820
発行日 / Pub. date	2007, 7

多段SVMを用いた頑健な動画ショット境界検出

宮村　祐一†　　中村　太一†　　篠田　浩一†　　古井貞熙†

† 東京工業大学 情報理工学研究科 計算工学専攻
東京都目黒区大岡山 2-12-1
E-mail: †{miyamura,nakamura}@ks.cs.titech.ac.jp, {shinoda,furui}@cs.titech.ac.jp

あらまし　本研究では、動画像ショット境界検出に複数のサポートベクターマシン (SVM) を 2 段階に構成して用いる手法を提案する。第 1 段階では、まず、ショット境界をその長さによって分類し、それぞれに対し異なる SVM を用いて認識を行う。さらに、明度の変化を用いたショット境界に対する専用の識別器を用いる。第 2 段階では、別の SVM を用いて誤識別の除去を行う。このような 2 段階処理により、ショット境界の種類によらない頑健な検出を実現する。TRECVID より提供された 13 本のニュース映像を用いて評価実験を行った。ショット境界検出の F 値は 84.5% となり、1 フレーム単位で認識を行う SVM のみの手法と比べて 3.3%改善し提案手法の有効性が確認された。また TRECVID2006 の結果と比較したところ、参加機関中で 3 番目に相当する認識性能を達成した。
キーワード　ショット境界検出、サポートベクターマシン、TRECVID、CBVIR

Shot Boundary Detection Using Multi-stage SVM

Yuichi MIYAMURA†, Taichi NAKAMURA†, Koichi SHINODA†, and Sadaoki FURUI†

† Department of Computer Science Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku Tokyo 152-8552 Japan
E-mail: †{miyamura,nakamura}@ks.cs.titech.ac.jp, {shinoda,furui}@cs.titech.ac.jp

Abstract　We propose a shot boundary detection using multi-stage SVMs. In the first stage, the shot boundaries are classified according to their length, and a SVM is constructed for each class. In addition, we use another calssifier to detect shot boundary with change in brightness. In the second stage, we remove the insertion errors by a SVM. This two-stage processing makes our system robust against a variety of shot boundaries. The proposed system was evaluated by using 13 news videos in TRECVID. The system reduced errors by 3.3% in F-measure from the conventional method using a frame-based SVM and achieved 84.5%. This result corresponds to the third recognition performance of TRECVID2006.
Key words　Shot Boundary Detection, Support Vector Machine, TRECVID, CBVIR

1. まえがき

近年、コンピュータ技術、特に、データの圧縮技術やネットワーク関連技術の発達を背景に、マルチメディアコンテンツの大量蓄積、高速転送が可能となり、テキスト、音声、静止画像と共に動画のコンテンツが増加している。特に、一般ユーザーが閲覧、蓄積できる動画データは、地上波ディジタル放送や HDD レコーダの普及に伴い、急激に増加している。ユーザーにとっては、手軽に大量のデータにアクセスできるようになる一方、大量のデータの中から本当に閲覧したい映像を捜し出すことが難しくなってきている。そこで、コンピュータによって自動的に動画コンテンツ対してラベルを付与する研究 (Content-Based Video Infomation Retrieval: CBVIR) が盛んに行われている。本研究では、この CBVIR の前処理として用いられるショット境界検出を目的とする。

ショット境界は境界の長さによって分類することができる。ショット境界の長さは、連続する 2 つのショット S_1、S_2 において、S_1 の最終フレームと S_2 の開始フレームの差として定義する。通常、境界の長さが 1 フレームのものを CUT、2 フレーム以上のものを GRADUAL と呼ぶ。CUT は、ショット S_1 の最終フレームの後ろにショット S_2 をただ繋ぎ合わせただけのものであり、境界前後のフレーム間で急激な変化を起こす最も単純なショット境界である。これに対し GRADUAL は、ショット S_1 の最後のフレーム画像とショット S_2 の最初のフレーム画像の間に 2 つのショットを滑らかに繋ぐためのフレームや、何らかのエフェクトのあるフレームが挿入されているものである。そのため、境界内の連続するフレーム間の変化は CUT と比べて少ない。また、各テレビ局、番組等が様々に違ったエフェク

(a) CUT

(b) GRADUAL

図 1　CUT と GRADUAL

トを入れていることもあり、その種類は無数に存在する。そのため、高精度な境界検出が困難となっている。

代表的なショット境界検出への取り組みとして、TREC Video Retrieval Evaluation(TRECVID) [1] と呼ばれる National Institute of Standards and Technology(NIST) が主催する動画検索を研究対象とするワークショップがある。そこでは、Distance Map を用いた手法 [2]、Finite State Machine（FSM）を用いた手法 [3]、サポートベクターマシン（SVM）を用いた手法 [4, 5] などが提案されている。特に、SVM を用いた手法は盛んに行われている。本研究では、複数の SVM を多段に組み合わせることにより、CUT だけでなく GRADUAL に対しても安定した性能をもつ頑健なショット境界検出手法を提案する。

本論文の構成は以下の通りである。続く 2 章で用いた特徴量を説明した後、3 章で提案する手法の説明をする。4 章で評価実験結果を報告し、5 章でまとめ、今後の課題を述べる。

2. 特　徴　量

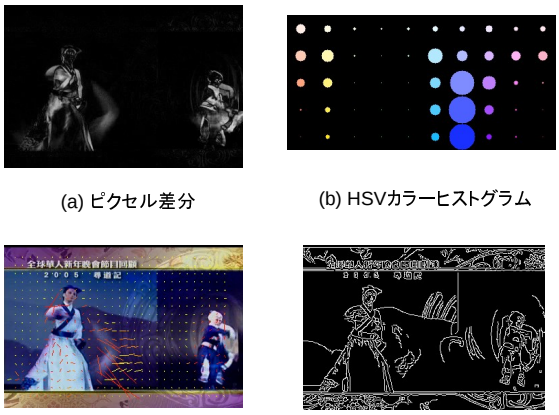
提案手法では、図 2 に示した 4 種類の低次特徴量を用いた。これらはすべて、ある 2 つのフレームの間の差分特徴量である。2 つのフレームをどのようにとるかについては、3.2 節で後述する。用いたフレーム画像は 352 × 240 ピクセルである。画像処理には OpenCV ライブラリ [6] を用いた。

2.1　ピクセル差分

ある 2 つのフレーム間の類似度として、フレーム画像同士の差分情報に注目する。対応するピクセルごとの画素値の差の絶対値をとり、それらを全てのピクセルで足し合わせた総和を求める。また、各ピクセルごとの画素値の差がある閾値を越えていれば 1、そうでなければ 0 とし、それらを全てのピクセルで足し合わせた総和を求める。ピクセル差分特徴量は、これらの差分の総和、閾値差分の総和の 2 次元となる。対象とするフレーム画像の間にショット境界や、被写体の動きがあればピクセル差分値は大きくなり、検出することができる。しかし、ピクセル差分のみによってショット境界であるのか、被写体の動き・カメラワークであるのかを判別することは困難である。

2.2　HSV カラーヒストグラム差分

フレーム画像をヒストグラム化することで、2 つのフレームの類似度を測る。カメラのフラッシュライトを誤ってショット境界と検出することを避けるために、色空間として HSV 色空間を採用する。V の値を除去することで HS の 2 次元となり、フラッシュライトによる明るさの変化に対し頑健な特徴量となる。HS の 2 次元色空間を 15 × 10 に分割し、与えられたフレーム画像の各ピクセルを各ビンに振り分ける。2 つのフレーム間で各ビンに含まれるピクセル数の差の絶対値をとり、全てのビンでそれらを足し合わせた総和を求める。また、各ビンに含まれるピクセル数の増加量、減少量の最大値も求める。HSV カラーヒストグラム差分特徴量は、これら変化量の総和、増加量の最大値、減少量の最大値の 3 次元となる。これらの特徴量の変化が大きければ、ショット境界や被写体の動きがあった可能性が高い。しかし、ピクセル差分と同様にショット境界と被写体の動きを分けること、また、色の変化の乏しいショット境界は検知することが困難である。



(a) ピクセル差分　(b) HSVカラーヒストグラム

(c) オプティカルフロー　(d)エッジ

図 2　特　徴　量

2.3　オプティカルフロー

オプティカルフローとは、画像撮影装置と被写体の相対的な動きをベクトルで表したものである。フレーム画像を格子状に 30 × 20 に分割し、各領域で動きベクトルを作成する。それらのベクトルの方向の分散値、平均ベクトルの大きさ、平均ベクトルの方向 (x、y 座標値)、追跡失敗数の値をとり、2 つのフレーム間で差をとることで 5 次元の特徴量を作成する。方向の分散値とは、共分散行列の 2 つの固有値の 2 乗和である。追跡

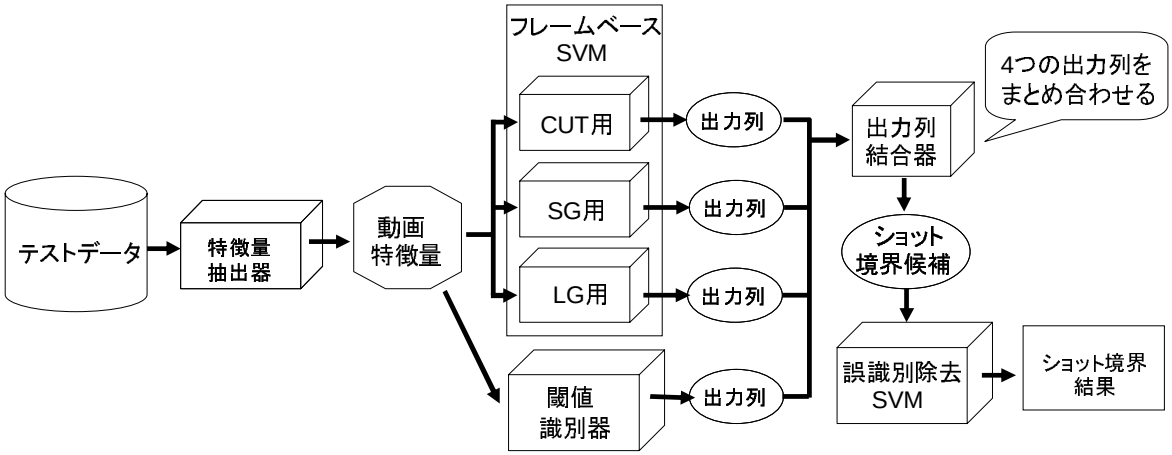


図 3　システムの概要

失敗数とは、オプティカルフローを計算するために被写体の動きを追跡するが、ショットの切り替わりなどで追跡ができなくなった領域の数を表す。オプティカルフローを用いることにより、被写体の動き、カメラワーク等を検出できる。これにより、ピクセル差分や HSV カラーヒストグラムでは困難なショット境界と被写体の動き、カメラワークの分離が可能となる。ここでは、オプティカルフローを求める手法として、Lucas-Kanade 法 [7] を用いる。

2.4　エッジ差分

エッジとは画像の中で画素値が急変する部分のことで、次のどちらかの条件を満たすものである。

（1） 画像の 1 階微分係数がある決められた閾値より大きい

（2） 画像の 2 階微分係数の符号が反転する

画像全体において、この条件を満たす画素数の和をエッジ量と定義する。これにより、2 つのフレーム間でエッジ量の増加量・減少量を求めることで 2 次元の特徴量が作成できる。ピクセル差分や HSV カラーヒストグラムは、色の変化の大きなショット境界では値が大きく変化するが、色の変化が乏しいときにはあまり有効ではない。エッジを用いることで、色の変化が乏しいときでも被写体の輪郭の違いにより画像の変化を検知することが出来るようになる。ここでは、エッジを求める手法として、Canny 法 [8] を用いる。

3. 提 案 手 法

3.1　システムの概要

本研究で提案するショット境界検出手法は、特徴量抽出器、出力列結合器と次の 3 種類の識別器によって構成されている。

（1） フレームベース SVM

（2） 閾値識別器

（3） 誤識別除去 SVM

システム概要を図 3 に示す。これらの識別器を用いて 2 段階処理を行う。第 1 段階として、テストデータの動画像から特徴量抽出器を用いて前章で述べた特徴量を抽出する。抽出された特徴量を 1 フレーム単位で認識を行うフレームベース SVM に与え、認識を行う。更にフレームベース SVM では検出が困難となる明度の単調増加・減少といった変化を用いたショット境界を検出するための専用の識別器を新たに追加する。このようなショット境界は SVM の学習を行うにはサンプル数が少ないため統計的な学習の有効性が少ない。しかしながら、閾値を用いた識別でも十分な認識性能が得られる。そこで、これらのショット境界に対しては SVM ではなく閾値を用いた識別器 (閾値識別器と呼ぶ) によって認識を行う。

次に、フレームベース SVM と閾値識別器からの認識結果をまとめ 1 つに統合し、ショット境界候補を作成する。境界候補は境界の開始フレーム番号と終了フレーム番号の組とする。

フレームベース SVM では、特徴量の差をとる際のフレーム間隔があらかじめ決められているため、その間隔より長いショット境界に対しては適切な認識が困難である。そこで、第 2 段階として誤識別除去 SVM を用いてショット境界候補ごとの認識を行う。これによって誤識別と認識された境界候補を除去し、残ったものが最終的なショット境界の結果となる。

3.2　フレームベース SVM

フレームベース SVM は、1 フレームごとにそのフレームがショット境界のフレームであるかどうか識別を行う SVM である。本システムでは、境界の長さに応じて以下の 3 種類を用意する。

（1） CUT 用 (ショット境界の長さが 1 フレーム)

（2） SHORT GRADUAL(SG) 用 (ショット境界の長さが 2〜5 フレーム)

（3） LONG GRADUAL(LG) 用 (ショット境界の長さが 6 フレーム以上)

SVM のカーネルとして CUT 用、SG 用 SVM では線形カーネルを、LG 用 SVM では RADIAL カーネルを用いる。また、2 章で述べた特徴量のうち実際に使用する特徴量の選択、差分計算に用いる 2 フレームの選択は、識別器により異なる。この特徴量とフレームの選択について、表 1 にまとめる。これらの

(a) 明度の変化を用いたショット境界

(b) フラッシュライト

図 4　フラッシュライト

表 1　フレームベース SVM で用いる特徴量。[x,y] は、フレーム t における特徴量を抽出する際に用いる 2 つのフレームの組。ここではあるフレーム t の特徴量について示している。

	特徴量		特徴量を抽出するフレーム画像の組
CUT	ピクセル差分	閾値差分の総和	[t-i-1,t-1][t-i,t][t-i+1,t+1], i=1,2
	オプティカル	x、y 座標値、大きさ	[t-4,t-1][t-3,t]
		追跡失敗数、分散値	[t-2,t-1][t-1,t][t,t+1]
	ヒストグラム	増加量、減少量	[t-1,t]
		変化量の総和	[t-1,t][t-2,t]
	エッジ差分	増加量、減少量	[t-2,t-1][t-1,t]
SG	ピクセル差分	閾値差分の総和	[t-i-1,t-1][t-i,t][t-i+1,t+1], i=1,2,4
	ヒストグラム	変化量の総和	[t-i-1,t-1][t-i,t][t-i+1,t+1], i=1,2
LG	ピクセル差分	ピクセル差分	[t-i-1,t-1][t-i,t][t-i+1,t+1], i=1,2,4,16
	オプティカル	大きさ	[t-4,t-1][t-3,t]
		分散値	[t-2,t-1][t-1,t][t,t+1]
	ヒストグラム	変化量の総和	[t-i-1,t-1][t-i,t][t-i+1,t+1], i=1,2,4
	エッジ差分	増加量、減少量	[t-2, t-1][t-1,t]

カーネルの種類、特徴量とフレームの選択は予備実験の結果に基づき決定した。

学習時には検出対象とするショット境界のフレームから抽出された特徴量ベクトルを正のサンプル、それ以外のフレームから抽出された特徴量ベクトルを負のサンプルとして SVM の学習を行う。

認識時には、テストデータ (動画像) の各フレームから作成される特徴量ベクトルを SVM に入力し、認識を行う。その認識結果により、各フレームは正・負の判別が行われる。正と識別されたフレームを 1、負と識別されたフレームを 0 と表し、1 番目のフレームから順に並べることにより 0、1 からなる列（出力列と呼ぶ）が作成できる。この処理を CUT 用、SG 用、LG 用の 3 つの SVM に対して行うことにより、3 本の出力列が作成される。出力列で 1 が連続しているときは、それらを 1 つのショット境界として扱い、先頭の 1 の前のフレームが境界の開始フレーム、最後尾の 1 のフレームが終了フレームとなる。LG 用の出力列では境界の長さが 6 以上のもののみをショット境界とする。

3.3　閾値識別器

明度の変化を用いたショット境界を検出するための識別器である。これらのショット境界は、次の 2 種類に分けられる。

（1）　ブラック境界：明度 V が単調減少することで、V が最小値付近まで達し、その後 V が単調増加することで次のショットに切り替わる (図 4(a) 上)。

（2）　ホワイト境界：明度 V が単調増加することで、V が最大値付近まで達し、その後 V が単調減少することで次のショットに切り替わる (図 4(a) 下)。

ここでは、ブラック境界の検出手法を説明する。なお、ホワイト境界に対しては逆の処理を行うことで検出可能である。$V(i)$、$R(i)$ を i フレーム目の明度の値、出力列 (閾値識別器の検出結果) の値とする。明度 V(i) の変化を表す式として $F(i)$ を次のように定義する。

$$F(i) = \frac{V(i+1) - V(i)}{V(i) - V(i-1)} \qquad (1)$$

$F(x) > 0$ のとき、$x-1$ フレーム目から $x+1$ フレーム目までの間は明度が単調増加・減少している。提案手法では、2 段階の閾値 H_s、h_s ($H_s < h_s$ とする) を用いることで検出を行う。出力列は次のアルゴリズムによって作成する。

（1）　全ての i において、$R(i) = 0$ とする。

（2）　全ての i において、次の 2 つの条件のうちどちらか一方でも成り立つとき、$R(i) = 1$ とする。

(a)　$V(i) < H_s$

(b)　$V(i) < h_s$ かつ $F(i) > 0$

（3）　全ての i において、$R(i+1) = 1$ もしくは $R(i-1) = 1$ で、$F(i) > 0$ を満たすとき、$R(i) = 1$ とする。

（4）　(3) を繰り返し、書き換えが起こらなくなったとき終了する。

表 2　フレームベース SVM による認識結果と提案手法による認識結果
R、P、F はそれぞれ Recall(%)、Precision(%)、Ｆ値 (%)

		フレーム	提案手法	差
ALL	R	82.2	83.4	+1.2
	P	80.3	85.6	+5.3
	F	81.2	84.5	+3.3
CUT	R	89.2	86.7	-2.5
	P	80.6	87.6	+7.0
	F	84.7	87.1	+2.4
GRAD	R	63.3	74.3	+11.0
	P	79.3	79.9	+0.6
	F	70.4	77.0	+6.6

（5） R(i)=1 となるフレーム列をショット境界とする。

なお、上記のアルゴリズム中の (2) の (b) と (3) では、フレーム画像の明度が単調増加・減少しているフレームを $F(i)$ を用いて検出しているが、提案手法では検出精度を高めるため、フレーム画像を 2 × 2 に分割し、全ての分割領域の明度が単調増加もしくは単調減少しているもののみを検出する。

3.4　出力列の結合

CUT 用、SG 用、LG 用のフレームベース SVM と閾値識別器によって作成された 4 つの出力列のショット境界検出結果を合わせて、1 つの結果にする。出力列同士の検出結果が競合した場合には、優先順位の高い方の結果を用いる。優先順位は、閾値識別器、CUT・SG 用、LG 用の順とする。CUT 用と SG 用の優先順位が同じ理由は、評価時に同じ種類のショット境界として扱われるため (4.1 節で述べる)、競合した場合を考慮する必要がないからである。このようにして結合された出力列の値が 1 のフレームがショット境界の候補となる。出力列で 1 が連続しているときは、3.2 節で述べたようにそれらを 1 つのショット境界として扱い、先頭の 1 の前のフレームが境界の開始フレーム、最後尾の 1 のフレームが終了フレームとなる。

3.5　誤識別除去 SVM

ショット境界の候補に対し、その開始フレームの画像と終了フレームの画像から特徴量を抽出する。開始フレームと終了フレームの特徴量の差をとり、ベクトル化して SVM に与え、認識を行う。フレームベース SVM ではあらかじめ決められた差分間隔で特徴量の差をとっていたが、誤識別除去 SVM ではフレームベース SVM と閾値識別器によって作成された境界候補の開始・終了フレームから特徴量を抽出し差をとるため、事前に差分間隔を設定する必要がなく、どのような長さのショット境界であってもその前後でショットが切り替わっているか判断可能となる。

4.　評 価 実 験

4.1　実 験 条 件

評価データとして、TRECVID より提供された CNN、LBC、CCTV 等のニュース番組を計 13 本用いた。総フレーム数は 597,043、ショット境界数は 3,785 である。

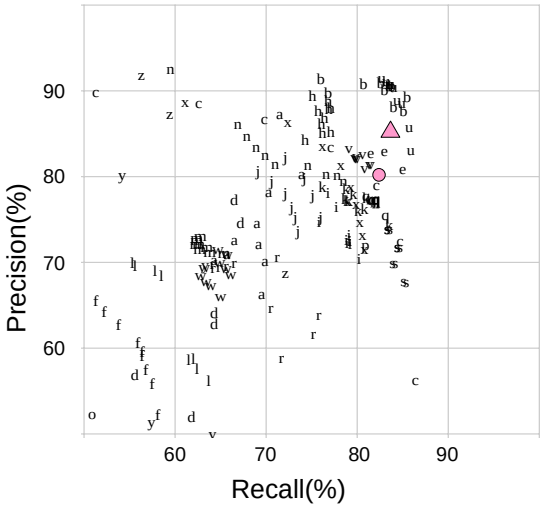


図 5　TRECVID2006 参加者の ALL の認識結果
△が提案手法、○が SVM のみ、アルファベットの点が各参加機関の結果を表す

提案手法で用いた SVM の学習時のパラメータ設定は、学習データを 2 つのグループに分け、一方を学習データ、もう一方をテストデータとして交差検定 (Cross Validation) を行い、結果の最も良いパラメータセットを用いた。また、TRECVID の方針により、ショット境界の長さが 5 以下のものを CUT、6 以上のものを GRADUAL として分類し、それぞれについて評価した。ALL とは全てのショット境界のことを指す。

4.2　評 価 方 法

評価尺度として、Precision と Recall を用いる。Precision(P) と Recall(R) は次のように計算される。

$$P[\%] = \frac{C}{S} \times 100 \ , \ R[\%] = \frac{C}{T} \times 100 \qquad (2)$$

C は認識結果に含まれるショット境界の数、S は認識結果の総数、T はショット境界の総数である。認識の目的、用途によって Precision または Recall のどちらか一方だけが重視される場合あるが、ほとんどの場合両方とも高い値になることが望ましい。Precision と Recall を同時に評価するために、F 値が使われている。F 値は Precision(P) と Recall(R) によって次のように定義される。

$$F = \frac{2PR}{P+R} \qquad (3)$$

4.3　実 験 結 果

提案手法を用いた認識実験を行った。また、提案手法の効果を確認するため、フレームベース SVM のみの手法でも評価実験を行った。その結果を表 2 に示す。さらに、TRECVID2006 の参加機関の認識結果を図 5、6、7 に示す。提案手法とフレームベース SVM のみの結果の比較を行うと、GRADUAL の Recall が 63.3%から 74.3%となり、提案手法を用いることで 11.0%ポイントの大幅な上昇があったことがわかる。これは、提案手法においてフレームベース SVM に加え、閾値識別器を追加したことによる効果と考えられる。また、第 2 段階において誤識別

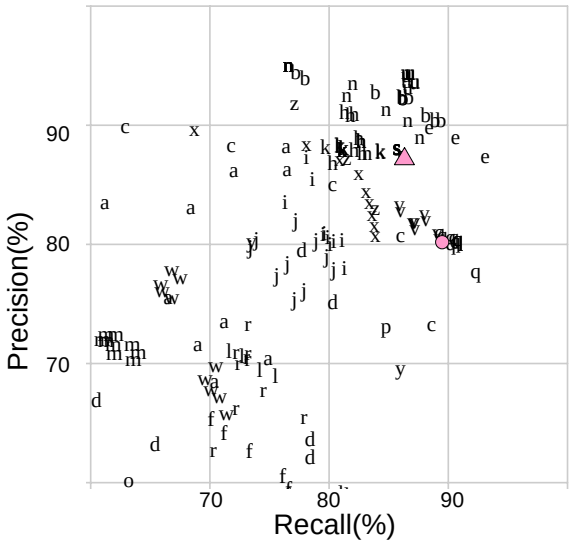


図 6　TRECVID2006 参加者の CUT の認識結果
△が提案手法、○が SVM のみ、アルファベットの点が各参加機関の結果を表す

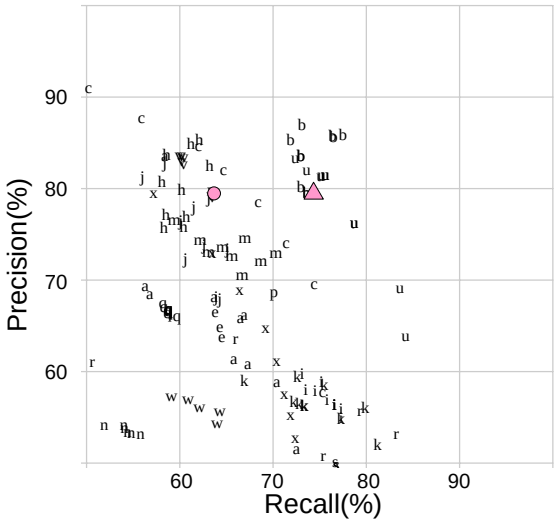


図 7　TRECVID2006 参加者の GRADUAL の認識結果
△が提案手法、○が SVM のみ、アルファベットの点が各参加機関の結果を表す

の除去をしたことにより、ALL、CUT、GRADUAL の Precision はそれぞれ 5.3%、7.0%、0.6%と全て上昇し、CUT では Recall が若干低下したものの、最終的に F 値はそれぞれ 3.3%、2.4%、6.6% 上昇した。TRECVID2006 の参加者の結果と比較を行うと、ALL の認識率は Recall が 83.4%、Precision が 85.6%となり全参加機関中で 3 番目の認識性能であることが分かる。これにより、提案手法の有効性が示された。

5.　まとめと今後の課題

本論文では、動画像ショット境界検出に複数の SVM を 2 段階に構成して用いる手法を提案した。第 1 段階では、ショット境界をその長さによって分類し、それぞれに対し異なるフレームベース SVM を用いた認識と、明度の変化を用いたショット境界に対しては専用の識別器を併用する。第 2 段階では、別の SVM を用いて誤識別の除去を行う。TRECVID より提供されたニュース映像を用いた評価実験を行ったところ、全てのショット境界に対しての認識性能は Recall が 83.4%、Precision が 85.6%、F 値が 84.5%となり、フレームベース SVM のみの手法に比べて F 値が 3.3%上昇した。TRECVID2006 の参加機関の認識結果と比較を行うと 3 番目の認識率であり、提案手法の有効性が確認された。

今後の課題として、次の 3 点が考えられる。まず、音声特徴量の利用が考えられる。提案手法では、画像から抽出した特徴量のみを用いたが、音声からの特徴量を併用することにより認識率を改善させることが可能だと考えられる。次に、本研究では、明度 V の変化を用いたショット境界に対し、SVM による認識とは別に新たな識別器を用いることで認識率の向上を行ったが、他種のショット境界においても、別の識別器を用いることで認識率のさらなる向上をさせることが可能だと考えられる。最後に、提案手法では、低次特徴量を複数用いたが、特徴量の数が多ければ多いほど動画像からの抽出に時間を費す。本研究では、複数の計算機を並列で実行させることによりその問題に対処したが、認識率を維持しつつ、計算量を削減する方法を検討したい。

文　　献

[1] P. O. et al.: “Trec 2005 video retrieval evaluation introductions”, Proc. of TRECVID2005 (2005).
[2] C. Cai, K. M. Lam and Z. Tan: “Trecvid2005 experiments in the hong kong polytechnic university: Shot boundary detection based on multi-step comparison scheme”, TRECVID 2005 Workshop (2005).
[3] D. G. Zhu Liu, E. Zavesky, B. Shahraray and P. Haffner: “At&t research at trecvid 2006”, TRECVID 2006 Workshop (2006).
[4] C.-W. Ngo, Z. Pan, X. Wei, X. Wu, H.-K. Tan and W. Zhao: “Motion driven approaches to shot boundary detection, low-level feature extraction and bbc rushes”, TRECVID 2005 Workshop (2005).
[5] C. Liu, H. Liu, S. Jiang, Q. H. Y. Zheng and W. Zhang: “Jdl at trecvid 2006 shot boundary detection”, TRECVID 2006 Workshop (2006).
[6] “Intel open source computer vision library”. http://www.intel.com/research/mrl/research/opencv/.
[7] B. Lucas and T. Kanade: “An iterative image registration technique with an application to stereo vision”, Proc. 7th International Joint Conference on Artificial Intelligence, pp. 674–679 (1981).
[8] J. Canny: “A computational approach to edge detection”, IEEE Trans. Pattern Analysis and Machine Intelligence,8, pp. 679–714 (1986).