カテゴリ化したフレーム情報における動画の自動シーン検出

佐藤　奏斗† 　南沢　樹† 　山岸　祐己† 　工藤　司†

† 静岡理工科大学情報学部　〒 437–0032 静岡県袋井市豊沢 2200–2

E-mail: †{2018048.sk,2018112.mi,yamagishi.yuki,kudo.tsukasa}@sist.ac.jp

あらまし　一般に，動画のフレーム情報として用いられる RGB 値や輝度，さらにそれらの差分などは，連続値として利用されるため，正規分布を仮定したモデルによる変化点検出が適当であるように思える．しかし，正規分布モデルは基本的に 1 次元情報の処理を想定しているため，多次元のフレーム情報を扱う場合は，各次元の計算結果を合算するなどの処理が必要となり，計算量も倍増する．よって，多次元のフレーム情報を扱うことを前提とした，高速な変化点検出手法の構築は重要であると言える．本手法は，カテゴリ化したフレーム情報の分布，すなわち多項分布を仮定しており，貪欲法に基づく決定的アルゴリズムによって，変化点数も自動で決定することを特徴とする．

キーワード　時系列データ処理，映像解析，汎用機械学習技術

1 はじめに

現在，動画のシーン検出は多様なアプローチが提案されているが，比較のためのベンチマークが確立されておらず，異なるデータセットのためのチューニングに労力が割かれているため，性能の比較が可能かつ自動で検出できる技術が求められている [1]．また，性能の向上が期待される技術として，集合知を利用するアプローチがあるものの [2] [3]，映像や音声といったデータの前処理が前提となっているため，大量の動画を高速に処理するためには，比較的小さいメタデータに対して適応できるものが望ましいと考えられる．例えば，単純なフレーム情報に適応可能な汎用的検出手法としては，MATLAB の findchangepts [4] が知られており，基本的には連続値の時系列データを扱うことを前提とした正規分布モデルが用いられている．しかし，この手法も，変化点数に関しては事前に設定する必要があり，時系列データの次元数とともに処理負荷が倍増する仕組みとなっている．よって，本論文では，多次元の情報を持ったフレーム情報を扱うことを前提とし，可能な限り単純化されたデータ，すなわち 1 次元化したカテゴリカルデータにおいて，自動的に変化点数を決定できるような手法を提案する．

2 提案手法

カテゴリカルデータ化した動画フレームデータに対し，多項分布を仮定したレジームスイッチング（変化点検出）手法を提案する．対称データを $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ とする．ここで，s_n と t_n は，J カテゴリの状態と n 番目のフレームをそれぞれ表す．$|\mathcal{D}| = N$ をフレーム数とすると，$t_1 \leqq \cdots \leqq t_n \leqq \cdots \leqq t_N$ となる．n はタイムステップとし，$\mathcal{N} = \{1, 2, \cdots, N\}$ をタイムステップ集合とする．また，k 番目のレジームの開始フレームを $T_k \in \mathcal{N}$ ，$\mathcal{T}_K = \{T_0, \cdots, T_k, \cdots, T_{K+1}\}$ をスイッチングタイムステップ集合とし，便宜上 $T_0 = 1$，$T_{K+1} = N + 1$ とする．すなわち，$T_1, \cdots, T_K$ は推定される個々のスイッチングタイムステップであり，$T_k < T_{k+1}$ を満たすとする．そして，$\mathcal{N}_k$ を k 番目のレジーム内のタイムステップ集合とし，各 $k \in \{0, \cdots, K\}$ に対して $\mathcal{N}_k = \{n \in \mathcal{N}; T_k \leqq n < T_{K+1}\}$ のように定義する．なお，$\mathcal{N} = \mathcal{N}_0 \cup \cdots \cup \mathcal{N}_K$ である．

いま，各レジームの状態分布が J カテゴリの多項分布に従うと仮定する，$\boldsymbol{p}_k$ を k 番目のレジームにおける多項分布の確率ベクトルとし，$\mathcal{P}_K$ はそれら確率ベクトルの集合，つまり $\mathcal{P}_K = \{\boldsymbol{p}_0, \cdots, \boldsymbol{p}_K\}$ とすると，$\mathcal{T}_K$ が与えられたときの対数尤度関数は以下のように定義できる．

$$L(\mathcal{D}; \mathcal{P}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log p_{k,j}. \tag{1}$$

ここで，$s_{n,j}$ は $s_n \in \{1, \cdots, J\}$ を

$$s_{n,j} = \begin{cases} 1 & \text{if } s_n = j; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

のように変換したダミー変数である．各レジーム $k = 0, \cdots, K$ と各状態 $j = 1, \cdots, J$ に対する式 (1) の最尤推定量は $\hat{p}_{k,j} = \sum_{n \in \mathcal{N}_k} s_{n,j} / |\mathcal{N}_k|$ のように与えられる．これらの推定量を式 (1) に代入すると以下の式が導ける．

$$L(\mathcal{D}; \hat{\mathcal{P}}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}. \tag{3}$$

したがって，スイッチングタイムステップの検出問題は，式 (3) を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

しかし，式 (3) だけでは $\mathcal{T}_K$ の導入によってどれだけ尤度が改善したかという直接的な評価をすることができない．この問題において，レジームスイッチングを考慮しないときの尤度からの改善度合いを評価することは重要であるため，尤度比最大化問題として目的関数を構築し直す．もし，レジームスイッチングのような変化が存在しない，すなわち $\mathcal{T}_0 = \emptyset$ と仮定すると，式 (3) は

$$L(\mathcal{D}; \hat{\mathcal{P}}_0, \mathcal{T}_0) = \sum_{n \in \mathcal{N}} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{0,j}, \tag{4}$$

となる．ここで，$\hat{p}_{0,j} = \sum_{n\in\mathcal{N}} s_{n,j}/N$ である．よって，K 個のスイッチングを持つ場合と，スイッチングを持たない場合の対数尤度比は

$$LR(\mathcal{T}_K) = L(\mathcal{D}; \hat{\mathcal{P}}_K, \mathcal{T}_K) - L(\mathcal{D}; \hat{\mathcal{P}}_0, \mathcal{T}_0). \tag{5}$$

のように与えられる．最終的に，この問題は上記の $LR(\mathcal{T}_K)$ を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

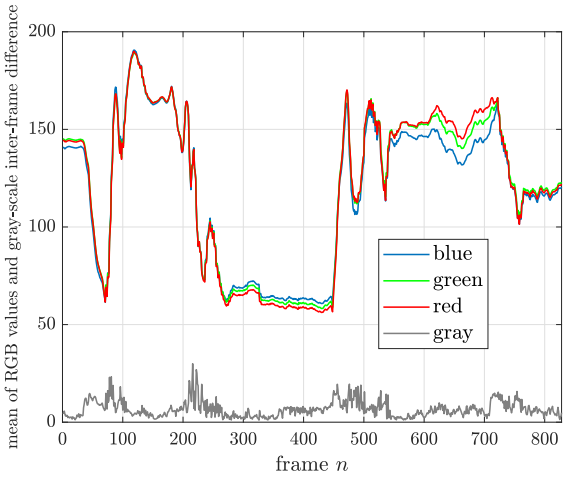
式 (5) を網羅的に解くと最適解が保証されるが，計算量が $O(N^K)$ となってしまうため，ある程度大きい N に対して $K \geqq 3$ となってしまうと，実用的な計算時間で解くことができない．したがって，任意の K について解くために，貪欲法と局所探索法を組み合わせた方法 [5] を用いる．なお，本実験では貪欲法アルゴリズムの終了条件として最小記述長原理 (MDL) [6] を採用し，事前にレジーム数，すなわち変化点数を設定することなく自動で終了させる．すなわち，このときの終了条件は下記となる．

$$-L(\mathcal{D}; \hat{\mathcal{P}}_k, \mathcal{T}_k) + \frac{(J-1)k\log N}{2} > -L(\mathcal{D}; \hat{\mathcal{P}}_{k-1}, \mathcal{T}_{k-1}) + \frac{(J-1)(k-1)\log N}{2}. \tag{6}$$

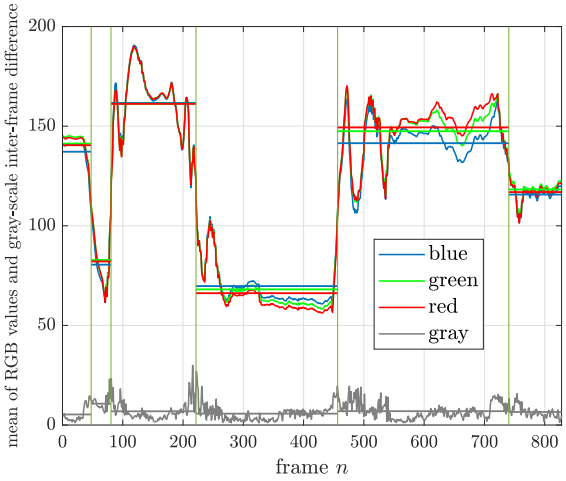
3　評価実験とまとめ

今回は，物体検出および文字認識を目的として撮影された短時間の動画で評価実験を行った．既存手法の findchangepts では，3 次元の色情報の平均値と，1 次元のグレースケールのフレーム間差分の平均値を使用し（合計 4 次元データ），提案手法では，さらにそれら各次元のフレーム間差分をとり，その最大値の次元をカテゴリ ($J = 4$) としたデータを使用した．ターゲットとなるシーンはフレーム 300 から 400 あたりの，対象文字にピントが合っているシーンだが，限られたフレーム数において，それ以外のシーンも検出できるかを確認する．提案手法で自動的にレジーム数が $K = 6$，すなわち変化点数が 5 となったため，既存手法の最大変化点数も 5 に設定した．

図 1, 2 より，既存手法と提案手法は，前半についてはおおむね同様の結果を検出しているが，後半については検出結果に差が出た．また，既存手法も提案手法も MATLAB R2022b で実行したが，フレーム数とカテゴリ数が共に少なかったこともあり，どちらも実行時間は平均して 0.025 秒ほどで差が出なかった．ただし，提案手法は，自動で変化点数を決定できているとともに，既存手法で利用しているデータ量（22.5kB）の約 55 分の 1 のデータ量（415B）でこれらを実現することができている．



(a) 色情報とグレースケールのフレーム間差分の平均値



(b) 最大変化点数を 5 に設定したときの結果

図 1: 既存手法 (findchangepts) によるシーン検出

文　献

[1] Manfred del Fabro and László Böszörményi. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia Systems*, Vol. 19, pp. 427–454, 2013.

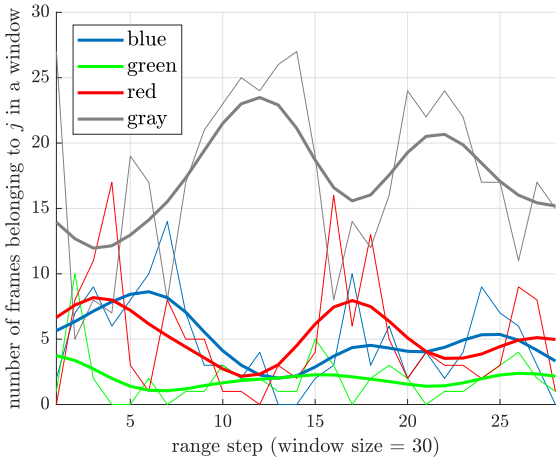
[2] Manfred Del Fabro and Laszlo Böszörmenyi. Summarization and presentation of real-life events using community-contributed content. In *Proceedings of the 18th International Conference on Advances in Multimedia Modeling*, pp. 630–632. Springer-Verlag, 2012.

[3] Wei-Ta Chu, Cheng-Jung Li, and Sheng-Chun Tseng. Travelmedia: An intelligent management system for media captured in travel. *Journal of Visual Communication and Image Representation*, Vol. 22, No. 1, pp. 93–104, 2011.

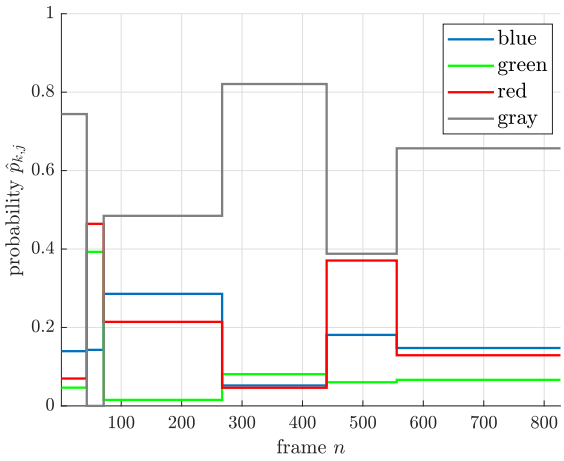
[4] Rebecca Killick, Paul Fearnhead, and I.A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, Vol. 107, pp. 1590–1598, 12 2012.

[5] Yuki Yamagishi and Kazumi Saito. Visualizing switching regimes based on multinomial distribution in buzz marketing sites. In *Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017*, Vol. 10352 of *Lecture Notes in Computer Science*, pp. 385–395. Springer, 2017.

[6] J. Rissanen. Modeling by shortest data description. *Automatica*, Vol. 14, No. 5, pp. 465–471, September 1978.



(a) 4 次元情報を 4 カテゴリ化したデータ



(b) 提案手法による自動検出結果 ($K = 6$)

図 2: 提案手法によるシーン検出