O-012

フレーム間差分情報を利用した動画の自動シーン検出

南沢 樹 †　　佐藤 奏斗 †　　山岸 祐己 †‡　　工藤 司 †
† 静岡理工科大学 情報学部 ‡ 理化学研究所 革新知能統合研究センター

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]，映像や音声といった大容量データの前処理が前提となっているため，エッジコンピューティング [3] を想定した大量の動画の高速処理を実現するためには，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，動画のフレーム情報として用いられる RGB 値やグレースケール値，さらにそれらの平均値や差分などは，正規分布を仮定した汎用的モデルによる変化点検出手法 [4] の適応が考えられる．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．実験では，カテゴリカルデータとしてフレーム間差分情報の最大値と最小値を扱う．

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，$t_1 \leq \cdots \leq t_n \leq \cdots \leq t_N$ 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 $\mathcal{N}_k = \{n \in \mathcal{N}; T_k \leq n < T_{K+1}\}$ 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

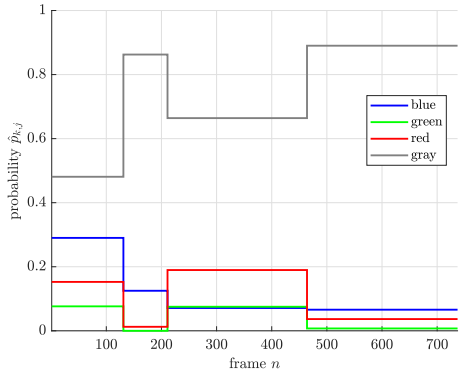
Automatic Scene Detection in Video Using Inter-frame Difference Information
†Ikki MINAMIZAWA　†Kanato SATO　†‡Yuki YAMAGISHI　†Tsukasa KUDO
†Shizuoka Institute of Science and Technology
‡RIKEN

に対して $K \geq 3$ 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．

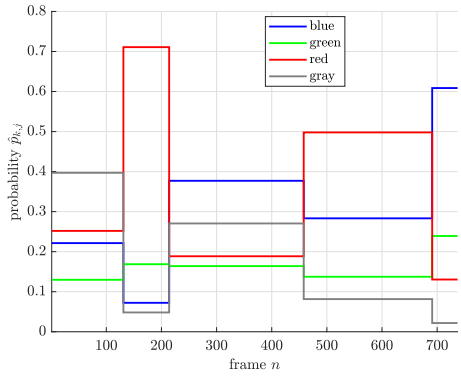
$$-L(\mathcal{D};\hat{\mathcal{P}}_k,\mathcal{T}_k) + \frac{(J-1)k\log N}{2} > -L(\mathcal{D};\hat{\mathcal{P}}_{k-1},\mathcal{T}_{k-1}) + \frac{(J-1)(k-1)\log N}{2}. \quad (6)$$

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．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），さらにそれら各次元のフレーム間差分をとり，その最大値もしくは最小値の次元をカテゴリ ($J = 4$) としたデータを使用した．図 1, 2 より，フレーム間差分最小値カテゴリは，フレーム間差分最大値カテゴリよりも，細かなシーンを自動で検出できていることが見て取れる．

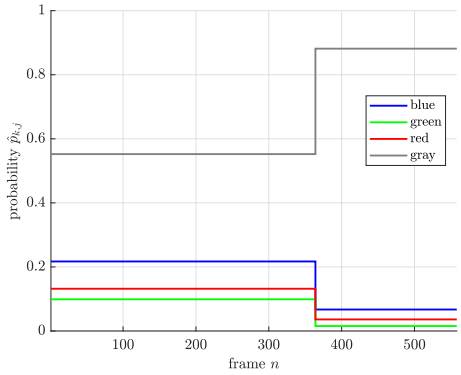


(a) フレーム間差分最大値カテゴリの結果

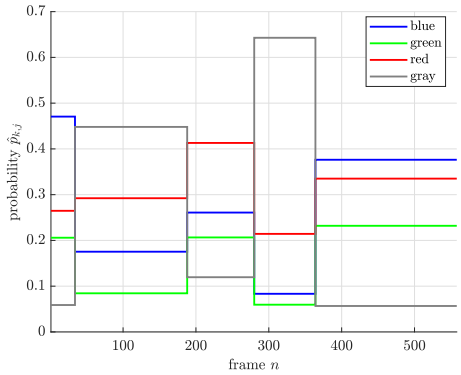


(b) フレーム間差分最小値カテゴリの結果

図 1: 動画データ 1 における提案手法の結果



(a) フレーム間差分最大値カテゴリの結果



(b) フレーム間差分最小値カテゴリの結果

図 2: 動画データ 2 における提案手法の結果

参考文献

[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] 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.

[3] Bin Liu, Zhongqiang Luo, Hongbo Chen, and Chengjie Li. A survey of state-of-the-art on edge computing: Theoretical models, technologies, directions, and development paths. *IEEE Access*, Vol. 10, pp. 54038–54063, 2022.

[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.