Earth Mover's Distance を用いた類似 Web 動画検索

高田 圭佑† 柳井 啓司†
電気通信大学 電気通信学部 情報工学科†

1 はじめに

近年，Web 上ではかつてないほど膨大な量の動画を閲覧できる状況になっている．しかしながら，それらの検索システムについては text-based なものが使用されているのが現状である．この場合，ユーザによって付加されたタグと呼ばれるキーワードを元に検索を行なうのが一般的であるが，付加するタグの選択はユーザの主観に依存するところが多く [1]，検索結果には様々な動画が存在してしまっている．

これらタグだけでは判別できない内容の差異を認識するためには，動画自身の特徴量を考慮することが重要であり [2]，これらの研究は，TRECVID [3] を中心に盛んに進められている．本研究では，これらの特徴量の比較基準に Earth Mover's Distance を利用し，キーフレーム数の異なる動画間での類似度算出を可能とすることで，キーフレームごとの特徴量を元にした類似動画検索を試みる．また，YouTube [4] から動画を収集し，評価実験を行なうことで，その有効性を示す．

2 方針

画像間の類似度を算出するための基準として，Rubner らによって提案された Earth Mover's Distance (EMD) [5] があるが，Peng らは，これを動画のクリップ間の類似度を算出するための基準として利用することで，ショット数が異なるクリップ間の類似度算出を可能とした [6]．本研究では，Peng らの手法を参考に，クリップ分割やキーフレーム抽出を簡略化する一方で，特徴量を追加するなどして改良を加えた手法を使用する．

3 検索手法

3.1 概要

本研究の検索手法の概要は，次のようになっている．

検索手法概要	
1：	YouTube から動画を収集する．
2：	動画からキーフレーム (シーン毎の中間フレーム) を抽出する．
3：	キーフレームを元に，特徴量を抽出する．
4：	特徴量から，各動画間の類似度を算出する．
5：	類似度を元に，検索結果を表示する．

3.2 特徴量抽出

本研究では，以下に示す 4 つの特徴量を利用する．なお，それぞれの値は $[0,1]$ になるように正規化を行なう．

i. 色特徴 キーフレームを 4 分割し，それぞれについてカラーヒストグラムを作成する．色空間は L*a*b* を利用し，次元数は $5 \times 3 \times 3 = 45$ 次元とする．

Web Video Retrieval Based on Earth Mover's Distance
†Keisuke Takada and Keiji Yanai, Department of Computer Science, Faculty of Electro-Communications, The University of Electro-Communications ({takada-k,yanai}@mm.cs.uec.ac.jp)

図 1 ランキング表示の例 (一番上がクエリ動画であり，以下類似度の降順に検索結果が並ぶ)

ii. 音特徴 キーフレームの前後合計 1 秒間の音声の大きさを解析し，それらの平均値を抽出する．

iii. オプティカルフロー キーフレームの前後合計 1 秒間について，全隣接フレーム間のオプティカルフローを算出し，それらの平均値を抽出する．オプティカルフローの算出には Lucas-Kanade 法 [7] を利用する．

iv. キーフレームポジション 動画の中でのキーフレームの再生位置を，始めを 0，終りを 1 として，$[0,1]$ の特徴量として抽出する．

3.3 類似度計算

抽出した特徴量を元に，全動画間の EMD に基づく類似度の算出を行なう．本研究では，キーフレームの特徴量でシグネチャを作成し，そのキーフレームが含まれるシーンの長さを各シグネチャの重みとすることで，EMD に基づいた類似度を算出する．類似度の値は $[0,1]$ で算出される．

3.4 結果表示

算出した類似度を元に類似動画の検索結果を表示する．本研究では，2 種類の表示方法を使用する．

i. ランキング表示 クエリ動画に対しての類似度を基準に，結果をランキング形式で表示する (図 1)．特徴量の重み付けは，ユーザが行なえるものとする．

ii. クラスタリング表示 検索対象の動画を，階層的クラスタリングを行なうことで複数のグループに分割し，動画を分類した状態で表示する (図 2)．本研究では，類似度の大きいものから順にグループ化していき，その類似度が閾値より小さくなるまで続けることでクラスタリングを行なう．グループ間の類似度の算出には群平均法 (group average method)，最短距離法 (nearest neighbor method)，最長距離法 (furthest neighbor method) の 3 つを利用し，ユーザが選択できるものとする．

図 2 クラスタリング表示の例 (各画像が各動画のサムネイルであり，動画数の降順に分類結果が並ぶ)

4 評価実験

ランキング表示とクラスタリング表示それぞれの方法について，評価実験を行なった．実験には「baseball」「basketball」「soccer」「tennis」「volleyball」の 5 つのタグでそれぞれ 100 個ずつ収集した合計 500 個 (総再生時間約 38 時間) の動画データベースを使用した．また，評価に用いた正解動画は，試合映像が含まれる動画の中で，類似していると人目で判断した動画であり，それぞれのタグから 10 個ずつ選択した．

4.1 ランキング表示

4.1.1 方法

各正解動画について，同一タグの残り全 99 個の動画に対してのランキング表示を行ない，それぞれのタグの上位 10 個までに対する平均適合率 (AP: average precision) と，その平均値 (MAP: mean average precision) を求める．求める値は，特徴量をそれぞれ単体で利用した場合と，タグごとに最適な重み付けで全特徴量を利用した場合の合計 5 つについてである．

4.1.2 結果

どのタグについても全特徴量を利用した場合が最も良く，最大は soccer で 0.56，MAP は 0.31 になった (図 3)．soccer においては，0.56 という数値を残しており，この結果においては，その有効性を示すことが出来たものと考えられる．

4.2 クラスタリング表示

4.2.1 方法

同一タグの動画をクラスタリングし，各グループの F 値 (F-measure) を求める．F 値とは，適合率 (precision) と再現率 (recall) の調和平均であり，

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (1)$$

の式で表される．本実験では，F 値の最も大きいグループを正解動画グループとし，その F 値を比較することによって，結果の評価を行なう．実験は，全特徴量を利用した場合の類似度を元に，閾値を 1.000 から 0.500 まで 0.025 刻みで変化させて行なった．

4.2.2 結果

最も結果の良かった，グループ間の類似度算出に郡平均法を用いた場合の結果と，対象の全動画を一度に表

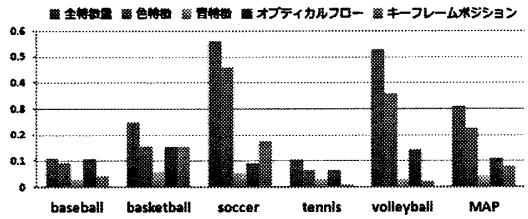

図 3 ランキング表示における平均適合率 (AP) とその平均値 (MAP)

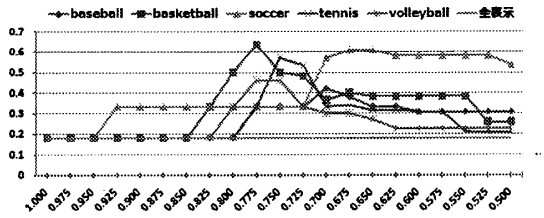

図 4 クラスタリング表示における各閾値での正解動画グループの F 値と，全表示した場合の F 値

示した場合 (全表示) の F 値を示す (図 4)．結果が最大となったのは basketball において閾値が 0.775 のときで，F 値は 0.64 となった．全表示の場合の F 値は 0.18 であり，クラスタリング表示を行なうことにより，類似動画を有効的に表示できたことが見てとれる．

5 考察

soccer と volleyball の結果が良かったのは，どの正解動画もコートの色が似通っていたためと考えられる．一方で，baseball や tennis の結果が悪かったのは，収集した動画に典型的な試合映像が少なく，正解動画自体の類似性に問題があったのではないかと思われる．

また，本研究の検索手法では，キーフレーム数 (カット点) の少ない動画において検索精度の低下が見られた．Web 動画においては，ユーザが撮影しただけの編集されていない動画も数多く存在するため，この点については，今後改善していく必要があるだろう．

6 まとめ

本研究では，EMD を用いた類似 Web 動画の検索手法について説明し，2 種類の表示方法を用いて実装を行なった．YouTube から収集した動画データベースに対する評価実験の結果，条件によっては，どちらの方法も有効であることが示された．今後は，新たな特徴量を追加していくと共に，キーフレームの抽出方法を改善することについても検討していきたい．

参考文献

[1] Yang, L., Liu, J., Yang, X. and Hua, X.: Multi-Modality Web Video Categorization, *Proc. of ACM MM-WS MIR*, pp. 265–274 (2007).

[2] 北研二, 津田和彦, 獅々堀正幹：情報検索アルゴリズム, 共立出版 (2002).

[3] TRECVID, http://www-nlpir.nist.gov/projects/trecvid/

[4] YouTube, http://www.youtube.com/

[5] Rubner, Y., Tomasi, C. and Guibas, L.: The Earth Mover's Distance as a Metric for Image Retrieval, *Int'l Journal of Computer Vision*, Vol. 40, No. 2, pp. 99–121 (2000).

[6] Peng, Y. and Ngo, C.: EMD-Based Video Clip Retrieval by Many-to-Many Matching, *CIVR 2005*, Vol. LNCS 3568, pp. 71–81 (2005).

[7] Lucas, B. and Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision, *Proc. of 7th IJCAI*, pp. 674–679 (1981).