論文 / 著書情報
Article / Book Information

論題	CNNから抽出した複数特徴量の統合に基づいた映像の意味インデクシング
著者	福田 竣, 井上 中順, 篠田 浩一
出典	第21回画像センシングシンポジウム (SSII) 講演論文集, , , IS2-16
発行日 / Issue date	2015, 6
Note	第21回画像センシングシンポジウム (SSII)講演論文集より転載

CNN から抽出した複数特徴量の統合に基づいた映像の意味インデクシング

福田 竣 † 井上 中順 † 篠田 浩一 †
† 東京工業大学
E-mail: {fukuda, inoue, shinoda}@ks.cs.titech.ac.jp

Abstract

映像コンテンツ解析の応用の一つに，意味のある対象 (コンセプト) が存在する映像を検出する意味インデクシングがある．このタスクに対して，映像ショットから複数フレームを取り出し，Convolutional Neural Network(CNN)[?] で抽出した特徴ベクトルを統合して識別を行う手法を提案する．またその統合方法として，特徴ベクトルの要素ごとに絶対値の最大値と平均偏角の複素数を算出する複素最大統合，及び特徴ベクトルごとに平均値を算出し，その値からの相対値で統合を行う平均基準法を提案する．評価実験として，TRECVID[?] の 2010 年度の映像コーパスから 30 のコンセプトを検出するタスクを行った．映像からフレーム画像 1 枚を抽出した時は MeanAP が 0.095 で，5 枚抽出した時は 0.123 となり提案手法の有用性が確認された．

1 目的・背景

近年記憶装置の大容量化や通信性能の向上により，インターネット上に多くの映像が存在している．その中から目的の映像を探し出すことが難しくなっている．そこで映像コンテンツ解析を用いて映像の検索性能を向上させることが必要である．

画像認識の手法の一つに Convolutional Neural Network(CNN)[?] がある．CNN は，大規模画像認識コンペティション ILSVRC2012[?] において，Krizhevsky ら [?] がエラー率を大きく下げたことで注目されている．ImageNet などで学習した CNN は汎用性が高く，他のデータセットに対しても，有用な特徴ベクトルを得ることができる [?].

そこで本研究では，映像から意味のある対象 (コンセプト) の検出を行う意味インデクシングに対して，ショットから複数フレームを取り出し，それらを ImageNet で学習済みの CNN で特徴抽出を行い，得られた特徴ベクトルを統合して識別を行う手法を提案する．

2 提案手法

ショットから複数のフレームを取り出し，ImageNet[?] でプリトレーニングした CNN でそれぞれ特徴抽出し，統合した特徴ベクトルを入力として SVM で識別を行う．

単純な統合方法として，複数の特徴ベクトルの平均を取る平均統合，要素ごとに絶対値が最大のものを選ぶ絶対値最大統合が挙げられる．本研究では新たに複素最大統合と平均基準法を提案する．

複素最大統合は要素ごとに絶対値の最大値と平均偏角の複素数を算出する手法である．複数の特徴ベクトルの要素 $x_1, x_2, \cdots, x_n$ を比較する際，$\mathbf{x} = (x_1, x_2, \cdots, x_n)$ とするとこれらの複素最大値 compmax($\mathbf{x}$) は次式で定義される．

$$\mathrm{compmax}(\mathbf{x}) = \max(|x_i|)\theta(\mathbf{x})$$

ここで，$\theta(\mathbf{x})$ は平均偏角であり，次式で定義される．

$$\theta(\mathbf{x}) = \frac{\sum_{i=1}^{n} |x_i| \mathrm{amp}(x_i)}{\sum_{j=1}^{n} |x_j|},\ \mathrm{amp}(x_i) = \begin{cases} 0\ (x_i > 0) \\ \frac{\pi}{2}\ (x_i = 0) \\ \pi\ (x_i < 0) \end{cases}$$

平均基準法は特徴ベクトルごとに平均値を算出し，その値からの相対値で統合を行う手法である．

3 評価実験

3.1 実験条件

映像解析のワークショップ TRECVID[?] の意味インデクシングのデータセットを用いて，30 種類のコンセプトの検出を行った．映像データは 2010 年度に用いられたインターネットビデオである．合計時間は学習用，テスト用ともに 200 時間である．ショット数は学習用が 119,685，テスト用が 144,998 である．

評価尺度は TRECVID で用いられている Mean Average Precision(MeanAP) を用いた．MeanAP はコンセプトごとの AP の平均である．AP は順位付き検索結果に対して次式で定義される．

$$\mathrm{AP} = \frac{1}{R} \sum_{r=1}^{L} \mathrm{Pr}(r)\mathrm{Rel}(r)$$

ここで，R は正解の総数，L はテストデータの総ショット数，$\mathrm{Pr}(r)$ は第 r 位までの Precision，$\mathrm{Rel}(r)$ は第 r 位の検索結果が正解であれば 1，不正解であれば 0 を出力する関数である．

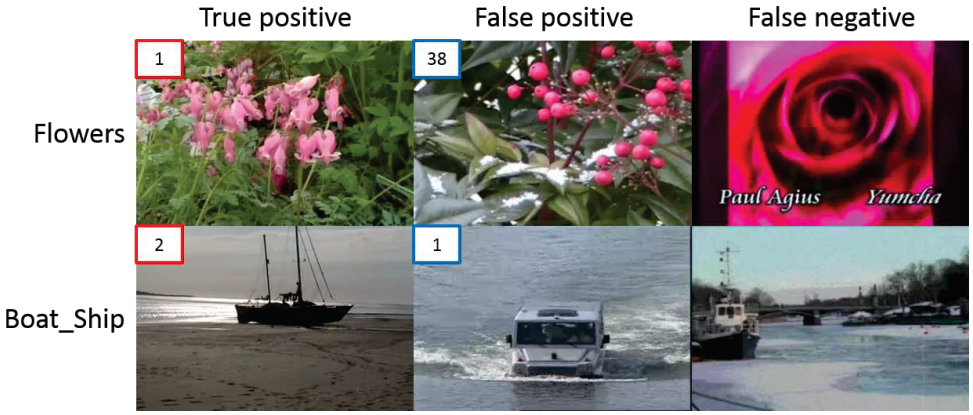

図 1　複素最大&平均基準で AP が低下したコンセプト例．左上の数は検索結果の順位である．

表 1　フレーム数に対する比較

フレーム数	MeanAP
1	0.0950
3	0.1126
5	0.1184
8	0.1216
12	0.1206

表 2　統合方法の比較

統合手法	MeanAP
平均	0.1120
絶対値最大	0.1184
複素最大	0.1205
絶対値最大&平均基準	0.1215
複素最大&平均基準	0.1232

3.2　実験結果

ショットから取り出すフレーム数を変えた場合と統合方法を変えた場合の実験を行った．

フレーム数に対する比較

絶対値最大統合を用いた時の，フレーム数変化に対する MeanAP の変化を表??に示す．有意水準 0.05 の検定を行ったところ，フレーム数が 5 までの増加に対して，AP が向上することは有意であることが確認された．

統合方法の比較

フレーム数が 5 で，統合方法変えた時の MeanAP を表??に示す．有意水準 0.05 の検定を行ったところ，絶対値最大&平均基準と複素最大&平均基準の組以外とは有意差が確認された．

具体的検出結果と誤り分析

複素最大&平均基準で AP が低下したものに，Flowers や Boat_Ship が挙げられる．これらの代表的な True positive，False positive，False negative を図??に示す．これらの図から背景の情報も特徴抽出する際に含まれていることが読み取れ，複素最大統合や平均基準法によりこの部分が大きく影響したことで AP が低下したと考えられる．

4　結論

本研究では，映像の意味インデクシングに対して，CNN を用いて複数フレームから特徴抽出し，統合した特徴ベクトルで識別を行う手法と，その統合方法として複素最大統合と平均基準法を提案した．フレーム数が 1 の時は MeanAP が 0.0950 であったが，5 の時は複素最大統合と平均基準法を用いて MeanAP が 0.1232 となった．

今後の課題として，Flowers や Boat_Ship などのコンセプトに対しては背景の情報を削るために，画像からコンセプトの範囲をトリミングした上で特徴抽出を行うことなどが挙げられる．

謝辞

本研究は JSPS 科研費 25280058 の助成を受けた．

参考文献

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks” *Proc. NIPS*, 2012.

[2] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVid,” *Proc. MIR*, pp.321-330, 2006.

[3] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database” *Proc. CVPR*, 2009.

[4] R. Girshick, J. Donahue, T. Darrell, J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” *Proc. CVPR*, 2014.