T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題	単語分散表現を用いた動画からのイベント検出
Title	
著者	金井怜, 井上中順, 李時旭, 篠田浩一
Author(s)	Satoshi Kanai, Nakamasa Inoue, Koichi Shinoda
出典	第21回 画像の認識・理解シンポジウム論文講演集, , ,
Citation	, , ,
発行日 / Pub. date	2018, 8

Powered by T2R2 (Tokyo Institute Research Repository)

CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

単語分散表現を用いた動画からのイベント検出

金井　怜[1,a)]　井上　中順[1,b)]　李　時旭[2,c)]　篠田　浩一[1,d)]

概要

物体，行動，場所などの複数の要素が関連している現実の出来事をイベントと定義し，動画からある特定のイベントを検出した．ここで，イベント毎に定義された文章や単語 (以下，イベント定義文) を物体，行動，場所などのカテゴリに分類し，カテゴリ毎の単語をフレームに映っている物体などのクラスと関連付けてイベントに対する重要なフレームを選択した．算出したスコアを mean Average Precision(mAP) で評価し，MED-14 Kindred データセットにおいて 34.6%から 36.4%に 1.8%の mAP の向上を，MED-17 PS(Pre-Specified) データセットにおいて 14.7%から 15.3%に 0.6%の mAP の向上を図ることができた．

1. はじめに

近年，YouTube や SNS などの動画投稿サービスにより，ウェブ上の動画は爆発的に増加する傾向にある．それに伴い，大量の動画から目的の動画を見つけるための技術が求められている．

まず，キーワードによる一般的な動画検索では，動画のタイトルや関連タグ以外の内容を検索することは困難を極める．そこで，動画を意味的に解析し検索するための試みとして画像認識技術が考えられるが，時系列情報を持つ動画において物体の変化に対応することができない．これに対し，物体の動きを認識する行動認識技術では，動作が重要ではない内容や他の物体との関連度，場所や音声といった情報を考慮する必要がある．

このように，動画検索のための認識技術は，物体，行動，場所，といった様々な情報が相互に関わるため，より複雑さを増している．同時に，動画検索だけではなく，監視カメラによる防犯対策や特定のイベントが起きた際に通知するシステムなどへの応用が考えられるため，より汎用的な手法が求められている．そこで，我々は，動画認識技術の発展が幅広く社会に貢献するその実用性を見込み，膨大な数の動画からある特定のイベントを見つけるための研究を行ってきた．

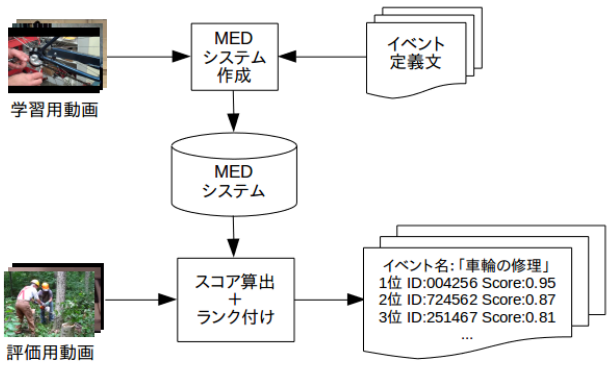

図 1　TRECVID MED タスクの概略図

これまでに，動画からのイベント検出に対して様々なアプローチが為されてきた．特に，画像を表す特徴表現の抽出に畳み込みニューラルネットワーク (Convolutional Neural Network：以下，CNN) が盛んに使用されている．CNN を用いた動画からのイベント検出の一般的な手法は，学習と評価の 2 段階に分けられ，次のような過程が挙げられる．1) 学習用動画のフレームから，事前学習済みの CNN を用いて画像特徴表現を抽出し，これを用いて識別器を学習させる．2) 評価用動画から抽出した画像特徴表現を，学習した識別器に入力しスコアを算出する．ここで，学習用動画が少ないことから効果的な画像特徴表現の学習が難しい，あるいはイベントに関係なく全てのフレームを同等に扱っている，といった課題が考えられる．

本稿では，単語を高次元の実数ベクトルで表現する単語分散表現を用いて，学習用動画の不足およびフレーム毎のイベントとの関連度を補う，Zero-Shot 手法について紹介する．イベント定義文の全ての単語を用いる場合，学習データを用いた単語毎の重み付けは，学習データに大きく依存することが考えられる．そこで，イベント定義文の内容をカテゴリに分割し，各々の単語分散表現を集約することで，学習データへの依存を抑制したカテゴリ毎の重み付けを行った．

1　東京工業大学
2　産業技術総合研究所
a)　kanai@ks.c.titech.ac.jp
b)　inoue@ks.cs.titech.ac.jp
c)　s.lee@aist.go.jp
d)　shinoda@c.titech.ac.jp

2. 関連研究

2.1 Multimedia Event Detection

Multimedia Event Detection は，大規模映像解析，映像検索に関する研究を促進するため，米国国立標準技術研究所 (NIST) が毎年開催している TRECVID(Text REtrieval Conference VIDeo retrieval evaluation)[1] と呼ばれる国際競争型プロジェクトのタスクの 1 つであり，これに世界中の多くの研究者グループが参加している．図 1 に TRECVID における Multimedia Event Detection タスク (以下，TRECVID MED タスク) の概略図を示す．TRECVID MED タスクは，与えられた少量の学習用動画とイベント定義文を基に，評価用動画からある特定のイベントを検出することを目的としている．これは，長さは数秒から数時間に渡り，イベントが複数の区間に含まれている，あるいはイベントが一切含まれていない多種多様な動画を対象としている．イベントには “家電の掃除” や “楽器の修理” などがあり，物体や行動が一意に定まらない複雑な要素で定義されている．

Zhao ら [2] は，フレーム毎に抽出した画像特徴表現を動画特徴表現に集約するため，VLAD (Vector of Locally Aggregated Descriptors) を使用した．これは，フレーム毎の画像特徴表現を平均する方法や Fisher Vector のような従来の特徴集約方法より，高速かつ簡潔に動画特徴として表現することができる．Sun ら [3] は，様々な構造の CNN から抽出した画像特徴表現のほか，物体特徴や音声特徴といった多種多様な特徴表現を試行錯誤的に組み合わせ，効果的なモデルを生成した．

これらの従来研究は，効果的な動画特徴表現を使用することでより良いスコアを得ることができるという考えに基づいている．しかし，効果的な動画特徴表現を抽出する際には，学習用動画の数が限られているという TRECVID MED タスクの問題点も同様に解決すべきであると考えられる．また，ウェブ上の動画は実際にイベントが発生している区間が明確ではないため，フレーム毎の画像特徴表現にも着目すべきであると考えられる．

2.2 Zero-Shot 学習

複数の要素で成り立っているイベント検出では，一般的に教師データのラベルを作成するために莫大な作業量を要し，学習用動画が少ない傾向にあるという課題を抱えている．そこで，教師データが少ないクラス，あるいは存在しない未知のクラスに対する予測が可能な手段として，Zero-Shot 学習 [4] が挙げられる．Sun ら [3] は，Word2Vec を用いてフレームに映っている物体などのクラスとイベント名を単語分散表現に変換し，これらの類似度を算出した．そして，類似度が低い，すなわちイベントとの関連度が低いフレームを除去することの有用性を示した．このように，TRECVID MED タスクにおいて，学習用動画が少ない点，あるいはイベント定義文を活用することができる点で，Zero-Shot 学習は適していると考えられる．

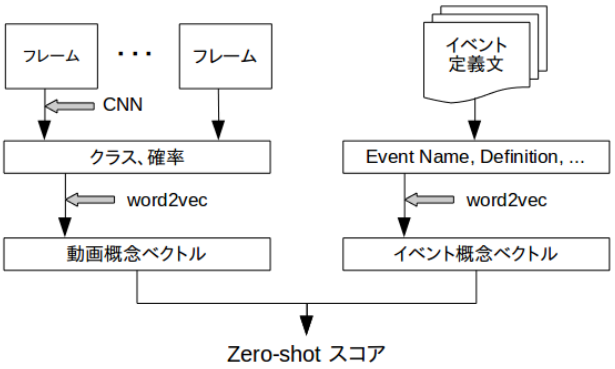

図 2 Zero-Shot 手法の概略図

3. 提案手法

本章では，提案手法である，単語分散表現を用いて動画からイベントを検出する Zero-Shot 手法について述べる．ある単語を数百次元のベクトルに変換し，単語のベクトル表現の獲得を可能にする Skip-Gram モデル [5] を用いて，イベント定義文の単語を単語分散表現に変換した．提案手法の概要図を図 2 に示す．

提案手法では，動画概念ベクトルとイベント概念ベクトル間のコサイン類似度を算出し，Zero-Shot スコアとして用いる．ここで，動画概念ベクトルは各動画内に映っている概念のクラスの単語分散表現を，イベント概念ベクトルは各イベント定義文の文章や単語から得られる単語分散表現を表す．

3.1 動画概念ベクトル

動画概念ベクトル $v(V)$ の定義を式 (1) に示す．

$$v(V) = \sum_{c} \left(\frac{1}{|n_V|} \sum_{i=1}^{n_V} p_{i,c} \right) \phi(c) \qquad (1)$$

ここで，i はフレーム番号，n_V は動画 V から抽出したフレーム数，$\phi(c)$ はクラス c の単語分散表現，$p_{i,c}$ は CNN から得られる事後確率を表す．動画概念ベクトル $v(V)$ は，次の 3 つの段階によって各動画 V から生成することができる．1) ImageNet[6] などのデータセットで学習済みの CNN を用いて，フレームから各クラス c の事後確率 $p_{i,c}$ を算出する．2) 各クラスを学習済みの Skip-Gram[5] モデルによって，単語分散表現に変換する．3) 式 (1) により動画概念ベクトルを得る．

3.2 イベント概念ベクトル

イベント E におけるイベント概念ベクトル $u(E)$ の定義

を式 (2) に示す．

$$u(E) = \sum_d \alpha_d \left(\frac{1}{|W_d(E)|} \sum_{w \in W_d(E)} \phi(w) \right) \tag{2}$$

ここで，d はイベント定義文のカテゴリ，$W_d(E)$ はイベント E のイベント定義文のカテゴリ d に含まれる単語の集合を，α_d は重みのパラメータ，$\phi(w)$ は単語 w の単語分散表現を表す．TRECVID MED タスクのイベント定義文は，“Event Name”, “Definition”, “Explication”, “Scene”, “Object/People”, “Activities”, “Audio” の 7 つのカテゴリを持つことに留意されたい．提案手法では，カテゴリの重み α_d を事前に学習するために，MED-14 Kindred データセットの学習用動画を用いて，式 (3) により各カテゴリ d の単語の集合から個別に mAP を算出した．

$$\alpha_d = \mathrm{mAP}_d \tag{3}$$

ここで，mAP_d は，カテゴリ d における $W_d(E)$ の単語セットのみを用いて算出した mAP を表す．提案手法の Zero-Shot スコア $s_{V,E}$ は，動画概念ベクトルとイベント概念ベクトル間のコサイン類似度を式 (4) により算出することで得られる．

$$s_{V,E} = \frac{v(V)^T u(E)}{\|v(V)\| \|u(E)\|} \tag{4}$$

4. 実験結果

4.1 実験条件

4.1.1 データセット

本実験では，学習用動画と評価用動画で構成された MED-14 Kindred データセット [7] と MED-17 PS データセット [8] を用いた．

MED-14 Kindred データセットは，20 種類（MED-17 Kindred データセットは 10 種類）の各イベントに 10 個の学習用動画が用意され，そのうちの 5 個が正例，5 個が負例に分割されている．さらに，全てのイベントで共通して用いられる 4992 個の負例の学習用動画と，システムの性能を測るために 12632 個（MED-17 Kindred データセットは 200000 個）の評価用動画が用意されている．

評価では，まず評価用動画から得られたスコアをイベント毎にスコアの高い順に並び替え，各々から平均適合率 (Average Precision : 以下，AP) を算出する．AP の定義を式 (5) に示す．

$$AP = \frac{1}{n} \sum_i (J_i \cdot P_i) \tag{5}$$

ここで，n は評価用動画の正例の数を，J_i は i 位の動画の正解判定値 (正例ならば 1 を，負例ならば 0 を取る) を，P_i は 1 位から i 位の動画における適合率 (Precision) を表す．イベント毎の AP の平均値を取った平均平均適合率 (mAP) を式 (6) に示す．

表 1　手法毎の mAP (%)

Run	MED-14 Kindred	MED-17 PS
SVM baseline	34.0	14.7
SVM + Zero-Shot	**36.4**	**15.3**

$$mAP = \frac{\sum_e AP_e}{E} \tag{6}$$

mAP の指標によって最終的な評価が為され，比較が行われる．ここで，AP_e はイベント e における AP，E はイベントの総数を表す．

4.1.2 モデル

本実験では，動画から 2 秒毎にフレームを抽出し，学習済みの CNN から求められる画像特徴表現として，GoogLeNet[9] の pool5/7x7_s1 層の出力を用いた．抽出した画像特徴表現は 1024 次元であり，これを入力としてサポートベクターマシーン (SVM)[10] により学習と評価を行った．また，GoogLeNet のソフトマックス層からの事後確率を提案手法に用いた．GoogLeNet は，12988 種類のクラスを持つ ImageNet-Shuffle[11] データセットで学習されたモデルを用いた．Skip-Gram モデルは，GoogleNews で学習されたモデルを用いた．これは，300 万個の語彙を対象に，各単語を 300 次元の単語分散表現に変換するモデルである．

4.2 実験結果

本実験の実行結果である手法毎の mAP を表 1 に，イベント毎の AP を図 3 に示す．ここで，“SVM baseline” は CNN と SVM で構成された識別器であり，“SVM + Zero-Shot” は SVM baseline 手法に Zero-Shot 識別器を追加した提案手法である．

表 1 より，MED-14 Kindred データセットにおいて 36.4%の mAP を，MED-17 PS データセットにおいて 15.3%の mAP を達成し，提案手法 (SVM + Zero-Shot) がベースライン手法 (SVM baseline) を上回る結果を得た．

実験結果より，提案手法の有効性が確認できた．しかし，図 3 に示した結果から，“ギター” や “ピアノ” といった複数の物体が現れる “楽器をチューニングする” のようなイベント以外には効果が低いことがわかる．

ベースライン手法では複数種類の物体が定義されていることによって少量の学習用動画がさらに細分化されるのに対し，提案手法では，図 4 より，“ギター” や “ピアノ” の単語分散表現間の類似度，あるいは “楽器” との単語分散表現間との類似度が非常に高いことが関係していると考えられる．すなわち，フレームに映っている物体などの形や色といった画像特徴表現による類似より，物体とイベント定義文が単語分散表現空間において近い距離にあることが性能向上の結果をもたらしたと考えられる．

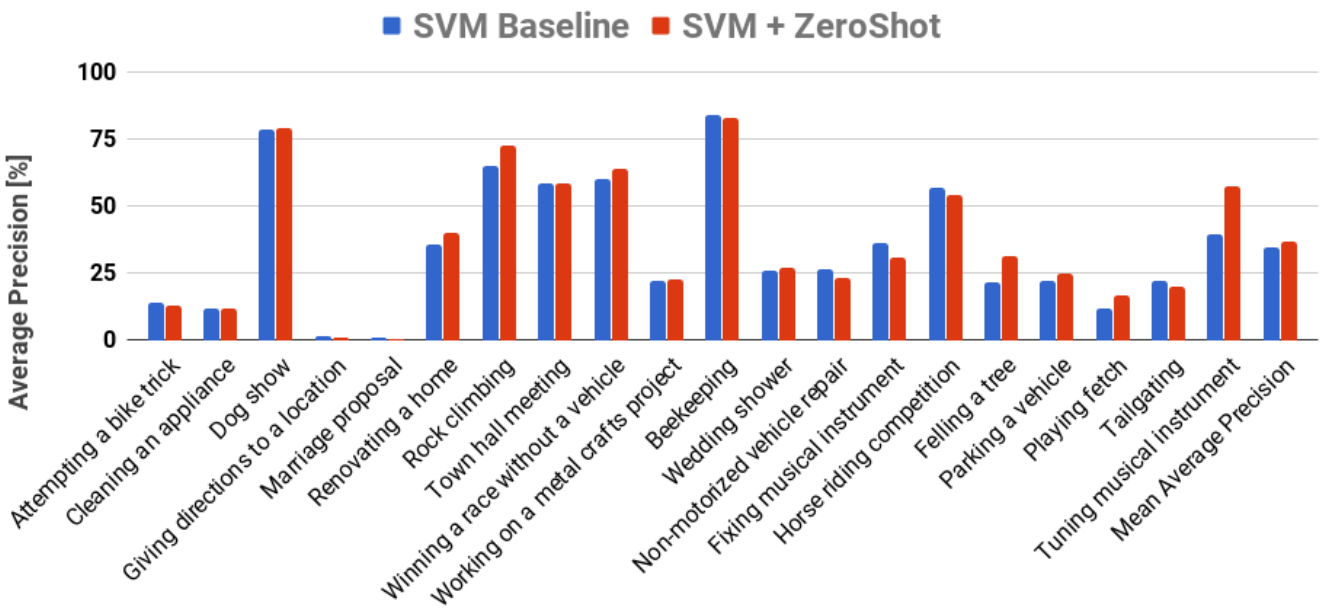

図 3　MED-14 Kindred データセットにおけるイベント毎の AP

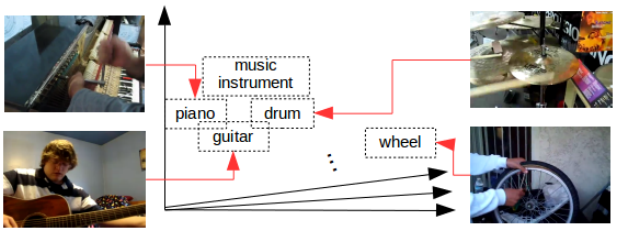

図 4　物体の word2vec 空間におけるコサイン類似度

5. 結論

本稿では，ウェブ上の増え続ける動画から目的の動画を探すことを想定した TRECVID MED タスクにおいて，学習用動画数が少ない点やイベント定義文を有効的に活用していない点などの問題を解決するため，イベント定義文のカテゴリ毎の単語分散表現を用いて得られた mAP を重みとし，荷重平均することによって単語分散表現を集約する手法を提案した．その結果，MED-14 Kindred データセットと MED-17 PS データセトにおいて，ベースライン手法の精度を上回る mAP の結果を得ることができた．今後は，時系列情報の消失や高精度なフレームへの重み付けに対応するために，LSTM(Long short-term memory) や注意機構を用いた手法について焦点を置きたい．

謝辞

本研究は，JST CREST JPMJCR1687 および JST ACT-I JPMJPR16V5 の支援を受けたものである．

参考文献

[1] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet, "TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking", in Proceedings of TRECVID 2017.

[2] Z. Zhao, M. Wang, R. Xiang, S. Zhao, K. Zhou, M. liu, S. He, Y. Zhu, Y. Zhao and F. Su, "BUPT-MCPRL team at TRECVID 2016: Multimedia Event Detection", in Proceedings of TRECVID, 2016.

[3] Y. Sun, R. Zhao, M. Li, C. Lu, H. Arai, T. Kinebuchi and Y. Jiang, "NTTFudan team at TRECVID 2016: Multimedia Event Detection", in Proceedings of TRECVID, 2016.

[4] M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell, "Zero-Shot Learning with Semantic Output Codes", in Advances in NIPS, pp. 1410–1418, 2009.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", in ICLR workshop, 2013.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", in Proceedings of CVPR, pp. 248–255, IEEE, 2009.

[7] "NIST. The Trecvid MED 2014 dataset", `https://www.nist.gov/itl/iad/mig/med-2014-evaluation`, 2014

[8] "NIST. The Trecvid MED 2017 dataset", `https://www.nist.gov/itl/iad/mig/med-2017-evaluation`, 2017

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", in Proceedings of CVPR, pp. 1–9, IEEE, 2015.

[10] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", in ACM transactions on intelligent systems and technology (TIST), 2011.

[11] P. Mettes, D. C. Koelma, and C. G. M. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection", in Proceedings of ICMR, pp. 175–182, ACM, 2016.