研究速報

物体検出器の検出傾向に基づく動画像のイベント検出

内海ゆづ子[†]（正員）　勝手　美紗[†]
岩村　雅一[†]（正員：シニア会員）
黄瀬　浩一[†]（正員：シニア会員）

Event Detection Based on Tendency of Object Detection

Yuzuko UTSUMI[†], *Member*, Misa KATTE[†], *Nonmember*, Masakazu IWAMURA[†], *Senior Member*, , , and Koichi KISE[†], *Senior Member*

[†] 大阪府立大学 大学院工学研究科, 〒 599–8531 大阪府堺市中区学園町 1–1
Graduate School of Engnieering, Osaka Prefecture University, 1–1, Gakuencho, Naka, Sakai, Osaka 599–8531, Japan

あらまし　動画像からイベントを検出する手法に，物体検出の結果を用いるものがある．これらは，動画像中からイベントに関係している情報のみ取り出すことで，精度よくイベント検出をしている．しかし，物体検出の誤りがあると，イベントの検出精度が低下する．一方で，類似した物体を誤検出しやすいなど，物体検出器の検出結果には傾向がある．そこで本論文では，物体検出器の検出の傾向を用いたイベント検出の可能性について検証する．

キーワード　イベント検出，物体検出，k 近傍法

1. はじめに

動画像を容易に撮影できるデバイスの普及に伴い，Web 上で動画像が増え続けている．動画像には，人の行動や催しなどのイベントが撮影されていることが多く，大量の動画像から特定のイベントを効率良く検索や分類をするためには，イベント検出の自動化が必要となる．

イベント検出には，大きく分けて 2 つの方法がある．1 つ目に，動画中の動きの情報を用いる方法が挙げられる [1]．あるイベントでは，動画中の物体がある特定の動きをするため，動きの情報はイベント検出に有用である．しかし，撮影方向の変化によって画像から得られる動きの特徴量が変化したり，カメラの動きが特徴量に反映されるため，動きの特徴量だけでは検出に失敗することがある．2 つ目に，イベントに登場する物体やイベントが発生する場所を用いる手法が挙げられる [2], [3]．イベントと物体や場所には強い相関があり，物体や場所の情報はイベント検出の有用な手がかりとなる．物体や場所を用いる手法では，動画像の撮影された場所や登場する物体の情報を数値化した Semantic Model Vector [2] や，動画像に登場する物体を用いる手法 [3] が提案されている．このうち後者 [3] は，画像全体から特徴抽出をする前者 [2] と比較して，動画像からイベントに関係する物体のみをイベント検出に用いることで，動画中のイベントに関係していない情報を排除できることから，より精度のよいイベント検出が可能と考えられる．本論文では，カメラの撮影方向が変化するものを含む動画像からイベントを検出することから，2 つ目に挙げた物体や場所を用いる方法でイベント検出を図る．

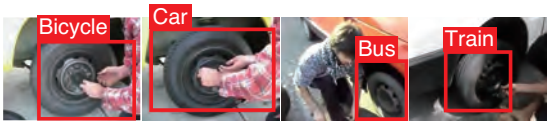

図 1　物体検出器 [4] の誤検出結果
Fig. 1　Examples of misdetection by using an object detector [4].

物体や場所を用いたイベント検出手法では，物体検出精度がイベント検出精度に影響を与える．しかし，誤検出しない物体検出器を作成するのは不可能と言っていい．1 つの方法は，識別器の検出の傾向をうまくイベント検出に役立てることであろう．図 1 は，物体検出器の 1 つである Felzenszwalb らの手法 [4] による物体検出の結果で，矩形は検出された物体の領域，ラベルは検出された物体を示している．結果から，タイヤをそれぞれ車，バス，自転車，電車など，タイヤや車輪がついている物体として検出しており，誤検出には傾向がある．この例の他に，検出対象とアピアランスが類似しているために誤検出が起こることもある [5]．これらから，物体検出器の検出傾向と画像の関係をあらかじめ学習しておけば，誤検出を起こす識別器を用いても，特定のイベントを判別できると考えられる．

そこで本論文では，物体の検出の傾向を表す特徴量を用いた動画像のイベント検出手法を提案し (注1)，動画像検索の手法を競うワークショップ TRECVID (注2) の Multimedia Event Detection (MED) task を通じて物体検出器の検出傾向を用いたイベント検出の可能性を検証する．

2. 提案手法の概要

提案手法の概要を図 2 に示す．提案手法ではまず，物体に関する特徴量を抽出するため，動画像からサンプリングした静止画に対して物体検出し，検出結果に

(注1)：提案手法は [9] に基づく．
(注2)：http://trecvid.nist.gov/

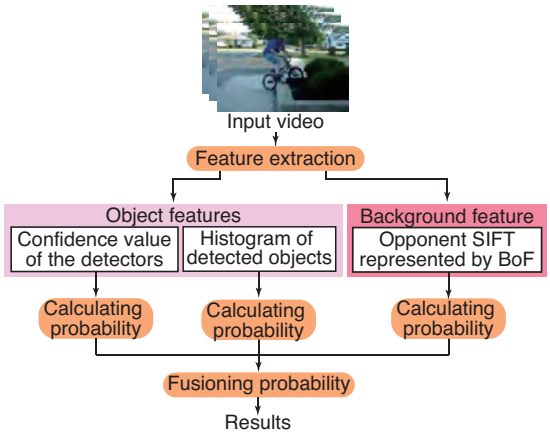

図 2　提案手法の概要
Fig. 2　Overview of the proposed method.

基づいて特徴量を算出する．イベントが行われる場所も物体と同様，重要な手掛かりとなるが，物体検出器では場所を検出できない．そこで，場所に関する特徴量を画像全体から抽出する．得られた特徴量それぞれで動画像中にイベントが含まれる確率を計算し，最終的に結果を統合する．以下，特徴量，確率の計算方法について詳細に述べる．

3. 特　徴　量

特徴量には，動画像中の物体の存在する可能性を表すだけでなく，検出の傾向を反映させるよう，物体検出器の出力値である信頼度と物体検出頻度を用いる．また，場所の情報を表すため，動画像からサンプリングした画像全体からも特徴を抽出する．以下，物体検出器に基づく特徴量，場所の特徴量の抽出方法について説明する．

3.1　物体検出器に基づく特徴量

まず，特徴抽出に用いる物体検出について説明する．イベントごとに関連した物体は異なるため，特徴抽出には複数の物体カテゴリを検出する必要がある．本論文では，それぞれ 1 つの物体カテゴリを検出する物体検出器を物体カテゴリの数だけ用意し，学習データを用いてあらかじめ学習しておき，各検出器を画像に適用して各物体を検出する． 物体検出は，各検出器で sliding window approach により探索窓を画像上で走査し，探索窓内に写っているものが検出対象の物体である確率を表す信頼度を計算する．そして，信頼度が閾値以上の探索窓を検出結果とする．この操作を物体カテゴリごとにして，得られた検出結果から特徴量を生成する．本論文では，Felzenszwalb らの提案する Deformable Part Model (DPM) を用いた物体検出器 [4] を用いる．

物体検出の結果から，2 種類の特徴量を抽出する．最初に，物体が存在する確率を表す特徴量について説明する．検出結果のうち，物体検出器の信頼度が大きいものほど，対象物体が検出された可能性が高い．そこで，文献 [11] に倣い，識別器の信頼度を物体の存在する可能性を表す特徴量として利用する． 選ばれた最大の信頼度は，並べて 1 つの特徴ベクトルとして表現する．このため，特徴ベクトルの次元数は検出する物体カテゴリの数に等しい．以降，この特徴量を信頼度特徴量と呼ぶことにする．

続いて，検出の傾向を表す特徴量について説明する．物体検出器の物体検出頻度は画像中の物体や検出する物体の種類によって変化する．そこで，物体カテゴリごとの検出の回数で検出の傾向を表す．この特徴量の特徴次元数は信頼度特徴量と同様，検出する物体カテゴリの数に等しい．以降，この特徴量を頻度特徴量と呼ぶ．

3.2　場所特徴量

イベントの中には，キッチンやバスルームなどの特定の場所で行われるものがあるため，これらの場所を区別できる特徴量は，イベント検出に有用となる．この場所の認識は一般物体認識と同じであることから，場所の表現をする特徴量として，一般物体認識で用いられる Bag-of-Features (BoF) model を用いる．提案手法では，局所特徴量に色情報を表す Opponent SIFT 特徴量 [6] を用いて，BoF Model に基づき特徴量の分布を表すヒストグラムを作成する．画像には背景だけでなく，物体が写っているが，一般的に物体と背景では背景の方が画像に占める割合が大きく，抽出された特徴量の大半は背景から抽出された特徴量であると考えられる．そこで，背景領域を推定せずに画像全体から特徴抽出をして，近似的に背景を表す特徴量とする．このヒストグラムを場所特徴量と呼ぶ．

4. イベント検出

イベント検出では，まず得られた 3 つの特徴量それぞれについて動画像中のイベント発生確率を計算する．そして，最終的に 3 つの特徴量から得られたイベント発生確率を統合することで，動画像中で特定のイベントが発生した確率を計算する．

特徴量ごとに特徴空間上で k 近傍をもとにイベント発生確率を計算する．いま，検出するイベン

トの総数を m，各イベントクラスを $\omega_1, \omega_2, \ldots, \omega_m$ とし，サンプル動画像の特徴量と正解クラスのセットを $s_i = (\boldsymbol{h}_i, \boldsymbol{c}_i, \boldsymbol{b}_i, \theta_i)$ ，サンプル動画像集合を $S = \{s_i | i = 1, 2, \ldots, n\}$ とする．ただし，$\boldsymbol{h}_i, \boldsymbol{c}_i, \boldsymbol{b}_i$ は物体検出の信頼度特徴量，頻度特徴量，場所特徴量を表し，θ_i は正解クラスで，$\omega_1, \omega_2, \ldots, \omega_m$ のいずれかを表し，n はサンプル動画像の総数を表す．クエリ動画像から得られた信頼度特徴量を $\boldsymbol{h}$, 特徴空間上での $\boldsymbol{h}$ の k 近傍を $N_{\boldsymbol{h}} = \{\boldsymbol{h}_{q_i} | i = 1, 2, \ldots k\}$ とする．ただし，$\boldsymbol{h}_{q_i}$ は $\boldsymbol{h}$ の i 番目の近傍特徴量を表し，$q_i \in \{1, 2, \ldots, n\}$ は $\boldsymbol{h}$ の k 近傍特徴量のインデックスとする．信頼度特徴量でイベント $\omega_s (s = 1, 2, \ldots, m)$ が発生する確率は

$$P(\omega_s | \boldsymbol{h}) = \frac{\sum_{r=1}^{k} \delta_{\theta_{q_r}, \omega_s}}{k} \tag{1}$$

と表される．ただし，$\delta_{\theta_{q_r}, \omega_s}$ はクロネッカーのデルタである．頻度特徴量，背景の特徴量でのイベント発生確率 $P(\omega_s | \boldsymbol{c}), P(\omega_s | \boldsymbol{b})$ も同様に計算する．最後に，それぞれの特徴量のイベント発生確率を統合してクエリ動画像のイベント発生確率を算出する．確率の統合の手法としては，さまざまなものが提案されている [10] が，予備実験の結果から，本論文では平均を用いて統合する．イベント発生確率をすべてのイベントで計算し，最もイベント発生確率が高いイベントをクエリ動画像に含まれているイベントとして検出する．

5. 実験と考察

提案手法のイベント検出精度を評価するため，実験を行った．提案手法の評価には，TRECVID2012 MED タスク [7] の DEV-O データセットを用いた．このデータセットは Web 上から収集した動画像から構成されており，各動画像の解像度や長さは異なっている．データセットのうち，learning set には 10 個の異なるイベントがあり，イベントごとに約 150 本の動画像が存在する．Test set には 31820 本の動画像がある．実験では，learning set を標本データ，test set を評価用のデータとして利用した．

物体の頻度特徴量と信頼度特徴量を抽出するのに，物体検出器を学習した．学習した物体は，“Aeroplane,” “Bicycle,” “Bird,” “Boat,” “Bottle,” “Bus,” “Car,” “Cat,” “Chair,” “Cow,” “Dining table,” “Dog,” “Horse,” “Motorbike,” “People,” “Potted plant,” “Sheep,” “Sofa,” “Train,” “TV/monitor,” “Person” の 21 種類である．これらの物体の学習には，PascalVOC2009 database [8] と INRIA Person Dataset (注3) を用いた．学習した物体の数が 21 のため，物体の信頼度特徴量と頻度特徴量の次元数はそれぞれ 21 となった．頻度特徴量を求めるための信頼度の閾値は，文献 [4] を実装した [12] に倣い，学習データを入力とした時に最も物体検出精度が高くなるように決定した．信頼度の閾値は各識別器で概ね -1 ± 0.5 となった．

頻度特徴量と信頼度特徴量は，1 つの動画像からランダムに 3 枚画像を抽出し，それらの画像における物体検出結果をもとに算出した．本実験では，標本データの動画像数が各カテゴリで 150 本と大きく，動画像から抽出する画像の枚数が少なくても物体検出の傾向が表せると考え，抽出画像を 3 枚に決定した． 場所の特徴量の計算には，動画像から 2 秒毎に抽出した画像を用いた．得られた Opponent SIFT 特徴量すべてを用いて 1 つのヒストグラムを生成した．BoF の数は予備実験の結果から 1800 とした．確率を算出するのに用いた k 近傍は，learning set の特徴量に対して 5-分割交差検証をすることで $k = 10$ に決定した．

イベント検出の評価基準として，イベントのカテゴリごとに検出結果の評価を行うために，次式で定義される Average Precision (AP) を用いた．

$$AP = \left(\frac{1}{N_R}\right) \sum_{l=1}^{N} \left(I_l \cdot \frac{N_{R_l}}{l}\right) \tag{2}$$

ただし，N はテストデータの総画像数，N_R はテストデータに含まれるイベント R の正解の画像数，N_{R_l} はイベント R を検出したときの上位 l 位内に順位付けされた動画像のうち，イベント R の動画像数であり，I_l は l 位に順位付けされた動画像がイベント R の場合は 1, そうでない場合は 0 となる．

5.1 実 験 結 果

1. で挙げた従来手法 [1]〜[3] のうち, [2], [3] は提案手法と公平な条件で比較できないため，物体の動きを用いる手法 [1] と提案手法を比較した. [2] は評価用動画像が本論文と違う上, [2] の特徴抽出に用いた識別器の学習データが非公開で再現実験ができず，比較できない．また, [3] は物体検出が成功すること前提とした手法で，評価でも物体検出が成功することを前提にしており，提案手法と公平に比較できない．提案手法と比較手法の各イベントの AP を表 1 に示す．

(注3)：N. Dalal, “INRIA person dataset,” http://pascal.inrialpes.fr/data/human/.

表 1　MED DEV-O dataset での AP
Table 1　AP of the proposed and comparative method with the MED DEV-O dataset.

Event Class	Tang et al. [1]	proposed method
Birthday party	**4.38%**	1.64%
Changing a vehicle tire	0.92%	**1.14%**
Flash mob gathering	**15.29%**	5.25%
Getting a vehicle unstuck	2.04%	**6.76%**
Grooming an animal	0.74%	**1.43%**
Making a sandwich	0.84%	**1.74%**
Parade	4.03%	**6.96%**
Parkour	**3.04%**	1.91%
Repairing an appliance	**10.88%**	10.65%
Working on a sewing project	**5.48%**	0.79%

表 1 から，イベント “Changing a vehicle tire”，“Getting a vehicle unstuck”，“Grooming an animal”，“Making a sandwich”，“Parade” で提案手法の方が従来手法と比較してよりよい結果を示した．これらの動画像では，イベント特有の物体が動画像に登場したり，特定の場所でイベントが行われていた．イベント特有の物体としては，“Changing a vehicle tire” や “Getting a vehicle unstuck” は車などのタイヤのある乗り物，“Grooming an animal” では猫や犬などの動物，“Making a sandwich” は食パンが挙げられる．また，イベントが行われた場所は，“Changing a vehicle tire” ではアスファルト舗装された道やガレージ，“Getting a vehicle unstuck” では地面，“Grooming an animal” は洗面台，“Making a sandwich” は台所，“Parade” は舗装道路であった．これらのイベント検出に成功したもののうち，“Grooming an animal” での物体検出結果の一部を図 3 に示す．図 3 から，“Cat”，“Dog”，“Cow”，“Horse” など四肢のある動物はお互いに誤検出をしていることがわかる．したがって，誤検出には傾向があり，誤検出結果に基づく特徴量はイベント特有の物体を表すことが可能である．従来手法の方が認識率が良くなった “Birthday party”，“Flash mob gahtering”，“Parkour”，“Repairing an appliance”，“Working on a sewing project” では，従来手法の方が提案手法よりも良い認識結果を示した．この原因として，イベントに関係する特定の物体が存在しなかったり，イベントが行われる場所が不特定であったことが挙げられる．物体や背景の情報だけでなく，イベントに特有の動きの情報や音声などの情報を用いることで，検出精度が向上する可能性がある．

6. ま と め

本論文では，動画像からのイベント検出を目的とし，物体検出器の検出結果の傾向を用いてイベントを検出する手法を提案した．正しく検出された物体を識別器の信頼度の最大値で，検出の傾向を物体検出器の検出頻度で，イベントが行われている場所を画像全体から得られる場所特徴量で表現した．これらの特徴量ごとに k 近傍法を用いてイベント発生確率を計算し，統合することで，最終的にイベントを検出した．実験の結果，特定の物体がイベントに登場する場合や特定の場所でイベントが行われる場合，提案手法は従来手法と比較してよい AP を示すことが明らかとなった．

今後の課題として，今回用いた物体などの情報だけでなく，動きや音の情報を用いることが挙げられる．

文　献

[1] K. Tang, L. Fei-Fei, and D. Koller, “Learning latent temporal structure for complex event detection,” Proc. of CVPR, pp.1250–1257, 2012.

[2] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, “Semantic model vectors for complex video event recognition,” IEEE Trans. on Multimedia, vol.14, no.1, pp.88–101, 2012.

[3] L. Jiang, A.G. Hauptmann, and G. Xiang, “Leveraging high-level and low-level features for multimedia event detection,” Proc. of the 20th ACM international conference on Multimedia, pp.449–458, 2012.

[4] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” Proc. of CVPR, pp.1–8, 2008.

[5] C. Hua, Y. Makihara, and Y. Yagi, “Pedestrian detection by using a spatio-temporal histogram of oriented gradients”, IEICE Trans. Inf. & Syst., vol. E96–D, no. 6, 2013.

[6] K.E. van deSande, T. Gevers, and C.G. Snoek, “Color descriptors for object category recognition,” Proc. of 4th European Conference on Colour in Graphics, Imaging, and Vision, pp.378–381, 2008.

[7] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A.F. Smeaton, and G. Quénot,

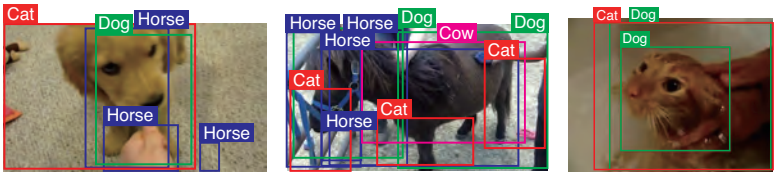

図 3 “Grooming an animal” での物体検出結果の例
Fig. 3 Examples of the detection results on “Grooming an animal”.

“TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics,” Proc. of TRECVID 2012, 2012.

[8] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, and A. Zisserman, “The pascal Visual Object Classes Challenge 2009 (VOC2009) Results,” http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/workshop/index.html, 2009.

[9] Y. Utsumi, M. Katte, M. Iwamura, and K. Kise, “Event Detection Based on Noisy Object Information,” Proc. of ACPR 2013, pp.572–573, 2013.

[10] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas “On Combining Classifiers,” IEEE Trans. on PAMI, Vol. 20, No. 3, pp. 226–239, 1998.

[11] M.A Sadeghi, A. Farhadi, “Recognition Using Visual Phrases,” Proc. on CVPR, pp. 1745–1752, 2011.

[12] R. B. Girshick, P. F. Felzenszwalb and D. McAllester, “Discriminatively Trained Deformable Part Models, Release 5,” http://people.cs.uchicago.edu/ rbg/latent-release5/.

(平成 xx 年 xx 月 xx 日受付)

Abstract Event detection methods from videos using results of object detection have been proposed. Detection errors of object detectors cause low accuracy of those event detection methods. However, there is a tendency of misdetection; object detectors misdetect objects which are similar to target objects. In this paper, we explore the possibility to detect events from videos with noisy object detection results. Experimental results suggested the possibility.

Key words Event detection, object detection, nearest neighbor search