þ,O ? ØCèô¶q
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

ô¶UC
TECHNICAL REPORT OF IEICE.

^hþ›0Åq`h°`ú.ÝÝwhŠwÌí CoHOG›Ã”wU|

¤ yœ—y Z±yG5y ô®y èyy ay°à y Iy8 y

y Êy0G¶ G¶ÃØCJ¶Z€J ß 464-8601 jŒ]Êy0¢• àÆÝÊ
yy 0Þ_™¶ G¶ &AØC¶æ ß 500-8288 0Þ]0Þ¢¤Û 1¸è 38j•
E-mail: ysnakamura@murase.m.is.nagoya-u.ac.jp, yyfddeguchi,ide,murase@is.nagoya-u.ac.jp,
yyttakahashi@gifu.shotoku.ac.jp

K'‡` Web ÍwG”w^hþ›®p‘XUgb”hŠwOAsAÉU[w°mq`o|^hþtéloM”ú.
›ÝÝb”U[U•['•”}HR|i-hþ›0Åq`h°`ú.ÝÝx]œtZ€Uæ˜•oVhU|^hþ›
0Åq`h°`ú.ÝÝwZ€x„qœræ˜•oMs M}^hþ›0Åq`h°`ú.ÝÝpx|^hþ¤w7‘
sÑè”ÜT'˜'•” Ý›Ãq^V›ÃwÒM›®L$tb;b”\qUOAqs”}ŠC px|2æ U
ZsrpôMQó›Ëm CoHOG¢Co-occurrence Histograms of Oriented Gradients£›Ì M²t¦Á`hÌí
CoHOG›Ã› Šb”}Ìí CoHOG›Ãx|^hþ¤wÁt–¬pwÌí ¯ M²wžIÎµÄ¬åÜpK
”}ÝÝtx| BoF¢Bag of Features£¯qq§”Éç SVM›;M”}îgpx| Web ÍT')B`h 10 §Â°æ|
- 1,000 ŠwˆhþŠ;M|¦ÓÂŸ§çÑé”›ÃS‘| SIFT ›Ãqz±b”\qt‘“|Ìí CoHOG›Ã
w ®Q›¬Ý`h}
©”ë”Å °`ú.ÝÝ|^hþ|Ìí CoHOG›Ã| BoF ¯q

A Study on Spatio-Temporal CoHOG Features for Recognition of Generic Objects in Video

Shogo NAKAMURA[y], Daisuke DEGUCHI[y], Tomokazu TAKAHASHI[yy],

Ichiro IDE[y], and Hiroshi MURASE[y]

y Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
yy Faculty of Economics and Information, Gifu Shotoku Gakuen University
Nakauzura 1{38, Gifu-shi, Gifu, 500-8288 Japan
E-mail: ysnakamura@murase.m.is.nagoya-u.ac.jp, yyfddeguchi,ide,murase@is.nagoya-u.ac.jp,
yyttakahashi@gifu.shotoku.ac.jp

Abstract Recognizing objects in videos is one of the important technologies to search a large amount of videos e -ciently on the Web. Recently, generic object recognition has been actively studied for still images, but almost not for videos. As for the generic object recognition in a video, it is important to use both the shape features and the motion features obtained from multiple frames in the video e ciently. In this paper, we propose spatio-temporal CoHOG (Co-occurrence Histograms of Oriented Gradients) features. This is an extension of CoHOG features that provide a high performance for pedestrian detection and others. The spatio-temporal CoHOG features are co-occurrence histograms of oriented spatio-temporal gradients in local regions in a video. In the recognition, a BoF (Bag of Features) representation and a kernel SVM are employed. We conducted an experiment on 1,000 videos including 10 categories collected from the Web. The experimental results showed the e ectiveness of the spatio-temporal CoHOG features compared with conventional optical ow features and SIFT features.

Key words Generic object recognition, Video, Spatio-temporal CoHOG features, BoF representation

1. x a Š t

Ùå| Web ÍtxG”w^hþU O`|f•'›®p‘X Ugb”U[U{Š'•oM”}Ugb”hŠw«¤æw°m q`o|^hþtq•”ú.U•['•"}^hþ¤wú.› ÝÝb”\qUpV•y|â”²x×üw_hM^hþ›0› ts`Zb\qUpV”}«Qy|×^ ›Ug«¤æqb” \qp|$ 1w‘Os^hþ›UgpV”}°`$t|\•' wUgx^hþtÇ5`h»¬¢Â©µÄ£›;Moæ˜•"} `T`sU'|\w‘Os»¬xâ”²U $tÇZ”‹w pK”hŠ|¯G••srUj¼pY`XUgpVsMÔùU K“|f‹f‹»¬UÇ5`sM‹w‹ M}fwhŠ|^h þ¤wú.›-‰;pÝÝb”U[UžAqs"}

Web Íw^hþt ‡•”ú.x7‘pK”hŠ|› wú .t‘ `sMÝÝ OUžApK”}îH„t ‡•”ú. ›-‰;U°`$sÊ¶pÝÝb” Ox°`ú.ÝÝqzy •|Ùå]œtZ€Uæ˜•oM” [1]}°`ú.ÝÝx§Â° æºp_hèwÌæ¤”³ãïUGVMú.›{OhŠ|ôM ÝÝp›˜”\qUÉ`X|r>b,V]JU M}

HR|i-hþ›0Åq`h°`ú.ÝÝ Ot b”Z €U]œtæ˜•oM”}E¯$s Oq`o| SIFT ¢Scale-Invariant Feature Transform £[2] ›;Mh‹wU•['•"} SIFT x|hþws8•°Ì!=t$HsÁt›ÃpK”\q UŒ'•oS“|¨Z`h›Ã:w*%–¬w¤¿´•Â«µ ½ßsrw Ý›Ã› 128 Íiw›ÃÕ«ÄçpG\b”‹w pK”}\•t0`o|^hþ›0Åq`h°`ú.ÝÝtS Mox Ý›ÃtCQo|^V›Ã›;M”\qU ®pK”}

Klaser 'x| HOG ¢Histograms of Oriented Gradients £[3] w¯ ›Ì M²t¦Á`h›Ã› Š`oM” [4]}\w›Ã x|^hþw ÝØCiZpsX¡€Ñè”Ü w^VØC‹ ˜”\qUpV|^^ÝÝt ®q^•oM”}°M|ó:w ›Ã”wžI›q”\qp0ÅwG\ó—›²Í^d”Z€U æ˜•oM”}f•'wZ€w°mq`o| Watanabe 'x¯ wžI›b;`h CoHOG ¢Co-occurrence Histograms of Oriented Gradients £› Š`oM” [5]}hþí ¤w7‘s •” p¯ wžI›q”\qpóvs ÝU¯qDóqs “|2æ UZt ®q^•oM”}

ŠC px|\•'w›ÃwQí›Êˆù˜dhÌí Co-HOG ›Ã› Šb”}é.$tx|HRw CoHOG tSZ” ¯ qžI›Ì M²t¦Áb”}fOb”\qp| Ýqˆ Vt0`oôMG\—›˜”}

Žñ| 2 ...pÌí CoHOG ›Ãqf•›;MhÝÝ Ot mMo\,"}f`o| 3 ...pîgwMOqAL›\,"} 4 ...px| 3 ...p\,hîgwßotCQo|Ìí CoHOG › Ãq w›Ã›wù`hÔùw®LtmMoÐ*`hAL›\ ,"}7™t| 5 ...pŠC ›‰b•}

2. Ìí CoHOG ›Ã›;MhÝÝ O

2.1 Ìí CoHOG ›Ã›;MhÝÝ Ow“A

$ 2 t| Owv•›Ôb}‡c|Ö—^hþT'o•à

$ 1 Web Íw^hþ

$ 2 Ìí CoHOG ›Ã›;MhÝÝ Owv•

›¨Z`|f\T'Ìí CoHOG ›Ã›¨Zb”}o• à x|È `hó:Ñè”ÜT's“|ÝÝ›æOo•p K”}f`o|o•à T'¨Z^•h›Ã› BoF ¢Bag of Features £[6] ¯q`|§Â°æ–t;™`h 2 «åµÝ +› ;Mo0Å§Â°æt b”ôTS›‰Zb”}7™t|¶o •à wôTST'|Ö—^hþw0Å§Â°æt b”ôT S›> b”}Žñp|¤^ŠtmMoÄ`X\,"}

2.2 Ìí CoHOG ›Ãw¨Z

o•à T'° ´pÌí Át–¬¢Òé¿«£›± ïÓæï¬`|Òé¿«–tÌí CoHOG ›Ã›¨Zb”} $ 3 t|Òé¿«w±ïÓæï¬w7 ›Ôb}Ž<pÌí CoHOG ›Ã”w-‰MOtmMo\,"}

‡c|Òé¿«°w¶hÉpÌí ¯ g(x;y;t) ›-‰b ”} I(x;y;t) ›hÉwKSq`o| g(x;y;t) w¤AÉ›Íw‘ Ot{Š"}

$$\begin{cases} g_x(x;y;t) = I(x+1;y;t) \quad I(x \quad 1;y;t) \\ g_y(x;y;t) = I(x;y+1;t) \quad I(x;y \quad 1;t) \\ g_t(x;y;t) = I(x;y;t+1) \quad I(x;y;t \quad 1) \end{cases} \quad (1)$$

Ít|Òé¿«›ó:w–¬¢·ç£tüÂ`|¤·çp É¯ g(x;y;t) ›-‰b”}f`o|¤ É¯ g(x;y;t) ›Y 20

$ 3 o•à T'wÒé¿«w±ïÓæï¬

$ 4 Y 20 Ø.¢$w:Èx" =M²›¯b£

Ø.wÖ:pK"p 12 M²t" =b"}Y 20 Ø.wÖ:w2 ªx| ' = $\frac{1+\sqrt{5}}{2}$ q`o (1; '; 0);(0; 1; ');('; 0; 1) pK"}$ 4 t|Y 20 Ø.›Ôb}

Òé¿«ºw¤·çp-‰`h¯ M²›b;`o|¯ M ²wžIÎµÄ¬åÜ›^Rb"}\•UÌí CoHOG ›Ã pK"}Ž<p|$ 5 t lo^Rwv•›†Ìb"}

¢a£ ‡c|,j·çqw•" t'"| 63 ¨wÖž› Rb"¢hi`|,j·ç× ‹ ‰£}sS| ² tÖž › R`sMwx|0¶Qt'"ÑÕQ› †b"hŠpK"}

¢b£ ¤Öžw•" tK"·çwÊ^ù˜d›Òé¿« ºT'»•`|fw¯ M²wÖž›ÎµÄ¬åÜtd®`o MX}

¢c£ ¤ÖžwÎµÄ¬åÜ›ÈAb"\qp74$sÎµ Ä¬åÜqb"}hi`|,j·ç‰œwÖžxfw·ç× w¯ M²w^pÎµÄ¬åÜ›^Rb"}‡h|¯ §SU 0ts`M·çxb;`sM}

Ìí CoHOG ›Ã"x (12 12 62) + 12 = 8,940 Íi qs"|‡×tÍiUGVXs"}fwhŠ| PCA ¢Principal Component Analysis £t'" dÍityVb"}

2.3 Ìí CoHOG ›Ãw BoF ¯q

ÝÝ›æO²rgq`o|o•à w¶Òé¿«p^R`h Ìí CoHOG ›Ã› BoF ¯qb"} BoF ¯qxhþ›Át ›Ã"wÎµÄ¬åÜp¯qb" OpK"|°`ú.ÝÝw üúp¿X;M'•oM"}\w Ox¶6^ŠqÝÝ^Št üT•oS"|°`tŽ<w qpæO}

¶6^Špx|‡c¤¶6;hþT'ó:wÁt›Ã"›¨ Zb"}f`o|¶¶6;hþw¶Át›Ã"› k-means « åµ»æï¬b"\qt'"| visual word ›\Rb"} visual word x|Át›Ã"›¯b›ÃÕ«Äç›Õ«Äç" =` h‹wpK"}¤Át›Ã"›7‹¨Åb" visual word q`

$ 5 Ìí CoHOG ›Ãw¨Z

o¯qb"\qt'"|¤¶6;hþx visual word wZqÃ SwÎµÄ¬åÜpG\^•"}ÝÝ^Špx|‡c¶6^Š q‰7t|¤Ö—;hþT'Át›Ã"›¨Zb"}f`o| ¶6^Šp\R^•h visual word ›;Mo|¤Ö—;hþ› visual word wZqÃSwÎµÄ¬åÜp¯qb"}

2.4 Ý +w¶6

¶6^hþwo•à p^R`hÌí CoHOG ›Ãw BoF ¯q›;Mo|Ý +›¶6b"}sS|Ý +x§Â°æ– t;™`oSX}Ý +tx|§"Éç SVM ¢Support Vector Machine £[7] ›;M"}§"Éç SVM xôMÝÝQó›ËI oM"\qUŒ'•|7'shþÝÝðJt ;^•oM"} ŠZ€px|§"Éç :q`o ² §"Éç›;M"} ² §"ÉçxŽ<wÜp¯^•"}

$$k(x;y) = \exp\left(\sum_i \frac{(x_i - y_i)^2}{x_i + y_i}\right) \quad (2)$$

² §"ÉçxhþÜ¨tSMo7‹QóU'M§"ÉçpK "qC ^•oM" [8]} tx|¶¶6Õ«ÄçwÊ^ù˜d tSZ" É ² 'mwo:›ƒb"}

2.5 o•à wÝÝ

Ä²t¶6`oSMhÝ +›;Mo|o•à wÝÝ›æ O}Ý +wZ—ALq`ox|ümÒ ØT'wÖøÇV' mU˜'•"}\w‹›0Å§Â°æt b"ôTSqb"}

2.6 w ù

o•à twôTS c_t ›;Mo|Ö—^hþwÝÝ›æO} Ö—^hþwo•à :› T qb"q|74$sôTS ¢xÍ Üt'lo> b"}

$$¢ = \frac{1}{T}\sum_{t=1}^{T} c_t \quad (3)$$

^hþtéloM"ú.›ÝÝ`hMÔù| ¢>0 qs"§ Â°æ›ÝÝALqb"}hi`|élоM"ú.U 1 ¨q MOÚEs'y| ¢U7Gw§Â°æ›ÝÝALqb"}‡h| «¤æ›)Qo^hþ›Ug`hMÔùx| ¢›fw‡‡b; b"}

(a) bicycle+person (b) bus

(c) car (d) cat

(e) cow (f) dog

(g) horse (h) motorbike+person

(i) person (j) sheep

$ 6 î g p – ; ` h ^ h þ w «

3. î g

3.1 î g Ã " » · ¿ Ä

Ã " » · ¿ Ä x | YouTube [9] T ' ^ h þ ›) B ` o Ï ™ ` h } ^ h þ ¤ w ° ` ú . w § Â ° æ x | PASCAL Visual Object Classes Challenge 2006 [10] þ – ; ^ • h Ã " » · ¿ Ä q ‰ ° w 10 § Â ° æ q ` h } é . $ t x | bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep p K " } $ 6 t Ã " » · ¿ Ä w ° æ › Ô b } Ã " » · ¿ Ä ¤ w ^ h þ x 0 Å ú . U é " ' O t ~ " Z ` h } ` T ` | ¦ « ç " ´ ã ï • h Ø Ž • w x ^ Z ` p ú . w ° æ U é l o M s M Ñ è " Ü U O ` h } ‡ h | Ÿ s " § Â ° æ w ú . U ‰ Ì t é l o M " ^ h þ ‹ ) B ` h } « Q y | bicycle, motorbike U ‡ • " ^ h þ t m M o x | b , o person ‹ ‰ Ì t é l o M " ‹ w p K l h } f w h Š | \ • ' w ^ h þ › person w Ý + t ¶ 6 ^ d " M x | b , o Ù ´ Â Ÿ Ò Ã " » q ` o { l h }

Ã " » · ¿ Ä w ^ h þ : x 1,000 Š p K " | ¤ § Â ° æ 100 Š ¢ person x bicycle, motorbike q ‰ Ì t é l o M " ^ h þ ‹ Š " q 300 Š £ q ` h } ‡ h | ^ h þ w Ñ è " Ü ± ¶ x 1,280 720 ™ 1,920 1,080 [pixels] | Ñ è " Ü : x 45 ™ 270 p K l h }

3.2 î g M O

Š b " Ì í CoHOG › Ã › ; M h O q | ¦ Ó Â Ÿ § ç Ñ é " › Ã | SIFT › Ã › ; M " Ž < w 2 m w O › z ± ` h } y z ± O 1 • Ý Ý O x | 2.2 ... › ¦ Ó Â Ÿ § ç Ñ é " › Ã t " V õ Q | 2.3 ... wBoF ¯ q › æ ˜ s M ‹ w q ` h } ¦ Ó Â Ÿ § ç Ñ é " › Ã q ` o x | ¡ € Ñ è " Ü w ¦ Ó Â Ÿ § ç Ñ é " › ; M h } Ñ é " w ¨ Z t x Ò é ¿ « Ú ¿ ½ ï ¬ O › b ; ` h } ¤ Ñ è " Ü w ¤ Ò é ¿ « t 0 ` o Ñ é " › - ‰ ` | o • à p Ñ é " w M ² Î µ Ä ¬ å Ü › ^ R ` h } y z ± O 2 • Ý Ý O x | 2.2 ... › SIFT › Ã t " V õ Q h ‹ w q ` h } SIFT › Ã : w U Z t x | DoG ¢ Di erence of Gaussian £ › ; M h sparse s O q ¬ æ ¿ Å ± ï Ó æ ï ¬ • å ï ¼ Ü ± ï Ó æ ï ¬ › ; M h dense s O U K " U | Š Z € p x ° ` ú . Ý Ý p ' " ® q ^ • " ™ w dense s O › ; M h } µ - " ç ! = t $ H s › Ã q b " h Š | › Ã : U Z x ó : µ - " ç w ¬ æ ¿ Å ± ï Ó æ ï ¬ t ' " æ l h }

¤ O › ; M " M w Í å Ý " » x | Í w ' O t ƒ ` h } Ì í CoHOG › Ã t m M o x | o • à w Ñ è " Ü : › f = 5 | Ò é ¿ « ± ¶ › $w_B = h_B = 80$ | · ç ± ¶ › $w_C = h_C = 5$, $f_C = 1$ | PCA & ; ™ w Í i : › $d = 500$ q ` h } ¦ Ó Â Ÿ § ç Ñ é " › Ã t m M o x | Ò é ¿ « ± ¶ › 20 20 | M ² Î µ Ä ¬ å Ü w Ï ï : › 36 q ` h } SIFT › Ã t m M o x | ± ï Ó æ ï ¬ ´ › 80 | µ - " ç › 8, 16, 24, 32 w 4 ¨ q ` h } Ì í CoHOG › Ã q SIFT › Ã w BoF t S Z " visual word w : x 500 q ` h }

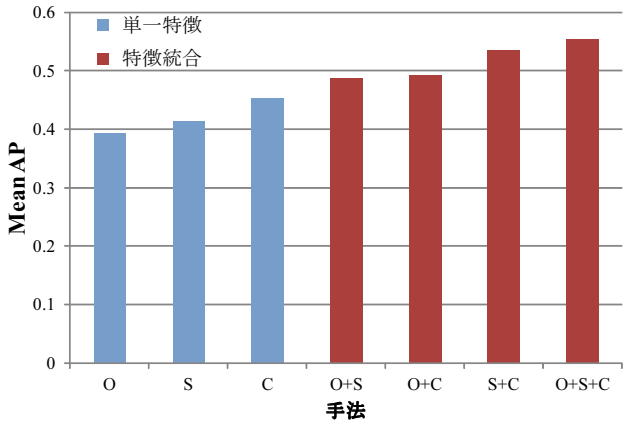

図 8　各特徴を用いた手法および特徴統合による手法の Mean AP（O はオプティカルフロー特徴，S は SIFT 特徴，C は時空間 CoHOG 特徴を表す）

で統合カーネルを作成する手法である．本来 MKL はカーネル選択のために作られたが，Varma らは MKL における各カーネルを特徴に対応付けて統合することで画像認識を行う手法を提案した [12]．そこで，本節でもカーネル選択ではなく特徴統合のために MKL を用いる．統合カーネル $K(\mathrm{x}, \mathrm{y})$ は特徴 f のカーネル $k_f(\mathrm{x}_f, \mathrm{y}_f)$ から次式のように作成できる．

$$K(\mathrm{x}, \mathrm{y}) = \sum_{f=1}^{F} \beta_f k_f(\mathrm{x}_f, \mathrm{y}_f) \tag{5}$$

F は特徴数，β_f は特徴 f の重みを表す（$\beta_f \geq 0, \sum_{f=1}^{F} \beta_f = 1$）．また，各特徴のカーネル $k_f(\mathrm{x}_f, \mathrm{y}_f)$ には，2 節同様 χ^2 カーネルを用いる．

4.2.2　特徴の統合実験

特徴統合の有効性を調査するため，各特徴のみを用いる手法と特徴統合を行う手法を比較した．実験条件は基本的に前節と同じであり，各特徴 f のカーネル重み β_f は均一重み（$\beta_f = \frac{1}{F}$）とした．

各特徴を用いた手法および特徴統合による手法の Mean AP を図 8 に示す．この図から，単一特徴を用いるよりも特徴統合を行う方が認識精度が高くなっていることがわかる．このことから，特徴統合の効果を確認した．また，特徴統合の中でも 3 種類すべての特徴を統合する手法が最も認識精度が良かった．すなわち，時空間 CoHOG 特徴は形状と動きの双方を記述するものの，SIFT 特徴とオプティカルフロー特徴には認識に有効な情報が他にあることがわかる．特に，オプティカルフロー特徴よりも SIFT 特徴と統合した方が認識精度が向上することから，時空間 CoHOG 特徴に形状情報が不足していると考えられる．そのため，空間方向をより重視した勾配の量子化が必要である．

5.　む　す　び

本報告では，動画像を対象とした一般物体認識のための時空間 CoHOG 特徴量を提案した．時空間 CoHOG 特徴は CoHOG における勾配と共起を時間方向に拡張したものである．単位区間から一定間隔で時空間 CoHOG 特徴を抽出し，PCA により次元を圧縮し，BoF 表現することで単位区間を記述した．そして，カーネル SVM による認識を単位区間毎に行い，それらの結果を平均することにより，入力動画像の認識を行った．実験では，YouTube において収集した 1,000 本の動画像を使用し，時空間 CoHOG 特徴とオプティカルフロー特徴，SIFT 特徴を比較した．オプティカルフロー特徴は，各隣接フレーム間のオプティカルフローのヒストグラムを作成することによって抽出した．SIFT 特徴は，各フレームからグリッドサンプリングにより特徴点を抽出し，BoF 表現することによって抽出した．実験の結果，時空間 CoHOG 特徴を用いた手法が最も高い認識精度となることを確認した．

今後の課題としては，より離れた時間間隔飴徳要椁引瓔産きリチ兒