

個人/グループサーベイ法

片岡 裕雄, *Ph.D.*

<http://www.hirokatsukataoka.net/>

片岡 裕雄（かたおか ひろかつ）

産業技術総合研究所 CV研究グループ/AL連携研究室/AIセンター 研究員
東京電機大学客員研究員



Hirokatsu KATAOKA, *Ph.D.*

片岡 裕雄

Research Scientist, AIST, Japan.

hirokatsu.kataoka [at] aist.go.jp

2014年慶應義塾大学大学院理工学研究科修了，博士（工学）。2013，2014年 ミュンヘン工科大学 Visiting Scientist，2014年東京大学博士研究員，2015年産総研特別研究員。2016年4月より現職。画像認識，動画解析，人物行動解析に興味を持つ。cvpaper.challenge主宰。2011年ViEW小田原賞，2013年電気学会誌論文奨励賞，2014年藤原賞，2016年ECCV WS Brave New Idea Award.

mypage: <http://www.hirokatsukataoka.net/>

cvpaper.challenge: <https://sites.google.com/site/cvpaperchallenge/>



大規模交通二アミス動画収集

AdaLEA: 時間的重みを操作した交通事故予測



文脈を弱教師とした異常検知



3次元畳み込みによる動画認識

片岡の主宰するcvpaper.challenge

論文読破・まとめ・発想・議論・実装・論文執筆・（社会実装）に至るまで取り組むCVの今を映す挑戦

- 人員：産総研，筑波大，東京電機大，慶應大による20名弱
- BraveNewなアイデアをトップ国際会議*に投稿

* Google Scholar Top-20にリストアップされている国際会議や論文誌

本取り組みの結果10本以上の論文（含CVPRx2, ICRA, CVPRWx5, ECCVWx2, ICCVW）が採択
8件の招待講演，3件の国内外での受賞



HP, Twitter, SlideShareもご覧ください

HP: <https://sites.google.com/site/cvpaperchallenge/>

Twitter: [@CVpaperChallenge](https://twitter.com/CVpaperChallenge)

SlideShare: [@cvpaperchallenge](https://slideshare.net/cvpaperchallenge)

サーベイ法

サーベイとは？

ひとことでいうと、その分野の動向を把握すること

- 現在どんな技術が流行っている？
- どういう歴史を辿ってきた？
- 自分のやっていることに最も近いものは？

II. RELATED WORK

A. Traffic data and approaches to its representation

Several practical databases for pedestrian detection, such as the French Institute for Research in Computer Science and Automation (INRIA) Dataset [3], Caltech [4], and the KITTI Vision Benchmark Suite for self-driving cars [2]) have been proposed in the past decade. The information contained in the KITTI database, which has been used to set meaningful vision problems for self-driving cars [2] as well as problems related to stereo vision, optical flow, visual odometry, semantic segmentation, two- and three-dimensional (2D/3D) object detection, and 2D/3D tracking, has proven especially useful.

In 2015, these problems were updated for stereo and optical flow [5]. Thanks to the development of sophisticated approaches, such as fully convolutional networks (FCN) [6] and region-based convolutional neural networks (R-CNN) [7], there has been improved performance of solving these problems using the KITTI benchmark dataset. In addition, a manner of geometry allows us to improve the rate of object detection [8] and optical flow [9] not only in stereo [10]. As for semantic segmentation, we can now obtain knowledge about dense connections and use this information with graphical models [11], [12].

Unfortunately, none of these datasets contain scenes of near-miss incidents in which pedestrians, cyclists, or other vehicles must be avoided. Thus, there is an urgent need for a collection of incident scenes that can be used to train self-driving cars on how to safely navigate such dangerous situations.

論文にもRelated workを書くこと多し

なぜ、サーベイをするのか？

トレンドの把握

- 知識がないと既存研究の劣化版を作りかねない
- トレンドを知らないと(天才でない限り)最先端の研究を生み出すことは難しい

自身の研究の立ち位置を確認

- 何が違う？なぜやる？どこが良いのか？という哲学

究極的には次のトレンドを作るため（ここ重要）

- 分野の方向性を自ら定める
- より良く、正しい方向へ導く

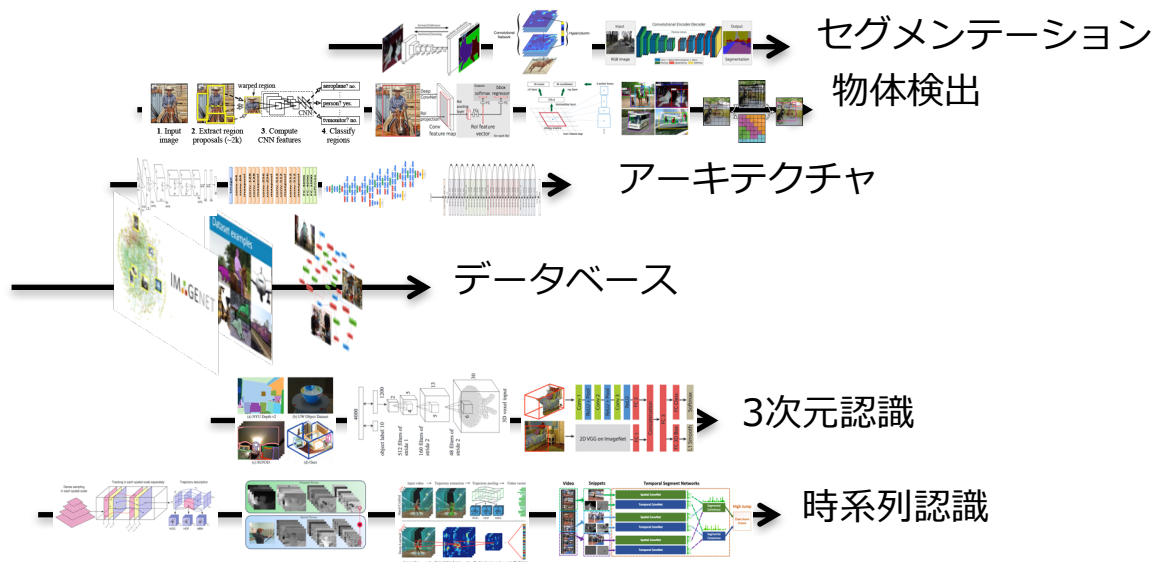
なぜ，サーベイするのか？

網羅的に調べることで分野の状況を精度よく映す

- サーベイの質や量が研究テーマ考案・選定に直結
- （グループ単位で）1,000本/年は論文を読むポテンシャルがあると良い

技術の流れがわかる

- 点より線/面で技術を捉える（参考にするなら歴史を追う，他分野から学ぶ）
- 次に何をすれば良いかがなんとなくわかる



どれくらいサーベイしてるの？

個人/グループとしてサーベイを推進

- 「個人」で達成
 - 2015年度 615本, 2016年度 400+本
- 「グループ」で達成
 - CVPR 2015 完全読破 (約10名)
 - 2016年, 1,000本読破達成 (約20名)

@2015

CVPR2015の論文

計

602

本を完全読破

@2016

論文読破

計

1,000

本を達成

意識的にサーベイして何が変わった？

2015年以前：個人プレー

- 論文調査：自分の狭い分野のみ
- 研究：従来法の単純な改善

組織的にサーベイしてから...

- 論文調査：網羅的に分野を把握
- 研究：物事の本質に迫るような問いを意識
- サーベイ/テーマ/実験に至るまで「質・効率を高める努力」を徹底
 - まだまだ徹底は不十分。。。

優れた問いを見つけよう！

新規に問題設定ができるようになった

- 思い付き（単純拡張）研究からの離脱
 - To invent, you need a good imagination and a pile of junk
 - 知識を詰め込むことで視野が繋がる感覚
- CVPR15完全読破から3年目, Top-tierの壁を突破
 - ICRA18採択！ (Robotics Top-1)
 - CVPR18採択x2！！ (CV&PR Top-1)

個人とグループの知識獲得がキモ

個人のサーベイ

グループのサーベイ

個人のサーベイ

グループのサーベイ

「分割と統合」の組み合わせが重要

個人サーベイで意識すること

速読と精読の組み合わせ

－ 速読（50+本/月）

- とにかく「広く浅く」
- 研究テーマやアイデア考案の時に行うサーベイ法

－ 精読（10-本/月）

- 実装レベルで「狭く深く」
- 具体的なテーマが決まっている際に

尺度を変えた読み方の統合で研究の効率化

気をつけていること

森を見る，木を見る

- － 森：ざっと全体を通して見る
 - 速読ならこれで十分
 - 背景，イントロ，図表，結果，結論を中心に
- － 木：細かいところまで目を通す（但し，目的を見失わない）
 - どんな情報が欲しいかを明確にして読む
 - 実装したい？ 輪講資料を作りたい？

森を見る (速読)

イントロ, 概念図を見よう

— どんな問題設定? その研究の新規性とは?

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)

Tsukuba, Ibaraki, Japan

{kensho.hara, hirokatsu.kataoka, yu.satou}@aist.go.jp

Abstract

The purpose of this study is to determine whether current video datasets have sufficient data for training very deep convolutional neural networks (CNNs) with spatio-temporal three-dimensional (3D) kernels. Recently, the performance levels of 3D CNNs in the field of action recognition have improved significantly. However, to date, conventional research has only explored relatively shallow 3D architectures. We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets. Based on the results of those experiments, the following conclusions could be obtained: (i) ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics. (ii) The Kinetics dataset has sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXt-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pre-trained simple 3D architectures outperforms complex 2D architectures, and the pretrained ResNeXt-101 achieved 94.5% and 70.2% on UCF-101 and HMDB-51, respectively. The use of 2D CNNs trained on ImageNet has produced significant progress in various tasks in image. We believe that using deep 3D CNNs together with Kinetics will retrace the success history of 2D CNNs and ImageNet, and stimulate advances in computer vision for videos. The codes and pretrained models used in this study are publicly available¹.

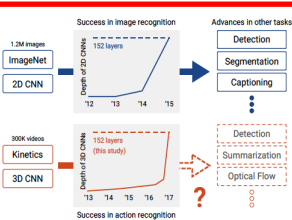


Figure 1: Recent advances in computer vision for images (top) and videos (bottom). The use of very deep 2D CNNs trained on ImageNet generates outstanding progress in image recognition as well as in various other tasks. Can the use of 3D CNNs trained on Kinetics generates similar progress in computer vision for videos?

more than a million images, has contributed substantially to the creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms, such as residual learning [10], have been used to improve image classification performance by adding increased depth to CNNs, and the use of very deep CNNs trained on ImageNet have facilitated the acquisition of generic feature representation. Using such feature representation, in turn, has significantly improved the performance of several other tasks including object detection, semantic segmentation, and image captioning (see top row in Figure 1).

To date, the video datasets available for action recognition have been relatively small when compared with image recognition datasets. Representative video datasets, such as UCF-101 [21] and HMDB-51 [17], can be used to provide realistic videos with sizes around 10 K, but even though they are still used as standard benchmarks, such datasets are obviously too small to be useful for optimizing CNN representations from scratch. In the last couple of years, ActivityNet [5], which

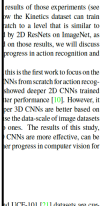
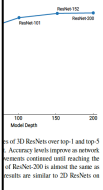


Figure 2: Success in image recognition. The use of 2D CNNs trained on ImageNet generates outstanding progress in image recognition as well as in various other tasks. Can the use of 3D CNNs trained on Kinetics generates similar progress in computer vision for videos?

including 250,000 action instances, ActivityNet also provides some additional tasks, such as unsupervised classification and detection, but the number of action instances is still on the order of tens of thousands. This year (2017), it was often cited to create a supervised pre-trained model. Kay et al. released the Kinetics dataset [14]. The Kinetics dataset includes more than 300,000 untrimmed videos covering 400 classes. In order to determine whether it can train deep 3D CNNs, we performed a number of experiments using these recent datasets, as well as the UCF-101 and HMDB-51 datasets.

Other large datasets, such as ActivityNet [13] and YouTube8M [1] have been proposed. Although these datasets are larger than Kinetics, their annotations are slightly sparser and only video-level labels have been provided. In other words, they include frames that do not relate to target actions. Such noise and the presence of unrelated frames have the potential to prevent these models from proceeding good training. In addition, with the size of sizes of 10 TB, their scales are simply too large to allow them to be utilized easily. Because of this, we will not refrain from discussing these datasets in this study.

2. Action Recognition Approaches

One of the popular approaches to CNN-based action recognition is to train a two-stream CNNs with 2D convolutional layers. In their study, Simonyan et al. proposed a method to use RGB and stacked optical flow frames as appearance and motion information, respectively [12], and showed that combining the two-streams has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-stream CNNs have been proposed to improve action recognition performance (e.g., [2, 22, 23]). Until the above-mentioned approaches, we focused on CNNs with 3D convolutional layers, which have been proposed in two-stream 3D CNNs through the use of large-scale video datasets. In this study, Simonyan et al. proposed a method to use RGB and stacked optical flow frames as appearance and motion information, respectively [12], and showed that combining the two-streams has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-stream CNNs have been proposed to improve action recognition performance (e.g., [2, 22, 23]). Until the above-mentioned approaches, we focused on CNNs with 3D convolutional layers, which have been proposed in two-stream 3D CNNs through the use of large-scale video datasets. In this study, Simonyan et al. proposed a method to use RGB and stacked optical flow frames as appearance and motion information, respectively [12], and showed that combining the two-streams has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-stream CNNs have been proposed to improve action recognition performance (e.g., [2, 22, 23]).

Method	UCF-101	HMDB-51
ResNeXt-18 (scratch)	42.4	17.1
ResNeXt-18	54.2	36.4
ResNeXt-34	57.7	39.1
ResNeXt-50	58.3	40.0
ResNeXt-101	62.8	43.3
ResNeXt-152	65.4	44.7
ResNeXt-201	68.1	46.5
ResNeXt-262	69.7	47.6
ResNeXt-312	70.7	48.6
ResNeXt-362	71.6	49.0
ResNeXt-412	72.4	49.4
ResNeXt-462	73.1	49.7
ResNeXt-512	73.8	50.0
ResNeXt-562	74.5	50.3
ResNeXt-612	75.2	50.6
ResNeXt-662	75.9	50.9
ResNeXt-712	76.6	51.2
ResNeXt-762	77.3	51.5
ResNeXt-812	78.0	51.8
ResNeXt-862	78.7	52.1
ResNeXt-912	79.4	52.4
ResNeXt-962	80.1	52.7
ResNeXt-1012	80.8	53.0
ResNeXt-1062	81.5	53.3
ResNeXt-1112	82.2	53.6
ResNeXt-1162	82.9	53.9
ResNeXt-1212	83.6	54.2
ResNeXt-1262	84.3	54.5
ResNeXt-1312	85.0	54.8
ResNeXt-1362	85.7	55.1
ResNeXt-1412	86.4	55.4
ResNeXt-1462	87.1	55.7
ResNeXt-1512	87.8	56.0
ResNeXt-1562	88.5	56.3
ResNeXt-1612	89.2	56.6
ResNeXt-1662	89.9	56.9
ResNeXt-1712	90.6	57.2
ResNeXt-1762	91.3	57.5
ResNeXt-1812	92.0	57.8
ResNeXt-1862	92.7	58.1
ResNeXt-1912	93.4	58.4
ResNeXt-1962	94.1	58.7
ResNeXt-2012	94.8	59.0
ResNeXt-2062	95.5	59.3
ResNeXt-2112	96.2	59.6
ResNeXt-2162	96.9	59.9
ResNeXt-2212	97.6	60.2
ResNeXt-2262	98.3	60.5
ResNeXt-2312	99.0	60.8
ResNeXt-2362	99.7	61.1
ResNeXt-2412	100.0	61.4
ResNeXt-2462	100.0	61.7
ResNeXt-2512	100.0	62.0
ResNeXt-2562	100.0	62.3
ResNeXt-2612	100.0	62.6
ResNeXt-2662	100.0	62.9
ResNeXt-2712	100.0	63.2
ResNeXt-2762	100.0	63.5
ResNeXt-2812	100.0	63.8
ResNeXt-2862	100.0	64.1
ResNeXt-2912	100.0	64.4
ResNeXt-2962	100.0	64.7
ResNeXt-3012	100.0	65.0
ResNeXt-3062	100.0	65.3
ResNeXt-3112	100.0	65.6
ResNeXt-3162	100.0	65.9
ResNeXt-3212	100.0	66.2
ResNeXt-3262	100.0	66.5
ResNeXt-3312	100.0	66.8
ResNeXt-3362	100.0	67.1
ResNeXt-3412	100.0	67.4
ResNeXt-3462	100.0	67.7
ResNeXt-3512	100.0	68.0
ResNeXt-3562	100.0	68.3
ResNeXt-3612	100.0	68.6
ResNeXt-3662	100.0	68.9
ResNeXt-3712	100.0	69.2
ResNeXt-3762	100.0	69.5
ResNeXt-3812	100.0	69.8
ResNeXt-3862	100.0	70.1
ResNeXt-3912	100.0	70.4
ResNeXt-3962	100.0	70.7
ResNeXt-4012	100.0	71.0
ResNeXt-4062	100.0	71.3
ResNeXt-4112	100.0	71.6
ResNeXt-4162	100.0	71.9
ResNeXt-4212	100.0	72.2
ResNeXt-4262	100.0	72.5
ResNeXt-4312	100.0	72.8
ResNeXt-4362	100.0	73.1
ResNeXt-4412	100.0	73.4
ResNeXt-4462	100.0	73.7
ResNeXt-4512	100.0	74.0
ResNeXt-4562	100.0	74.3
ResNeXt-4612	100.0	74.6
ResNeXt-4662	100.0	74.9
ResNeXt-4712	100.0	75.2
ResNeXt-4762	100.0	75.5
ResNeXt-4812	100.0	75.8
ResNeXt-4862	100.0	76.1
ResNeXt-4912	100.0	76.4
ResNeXt-4962	100.0	76.7
ResNeXt-5012	100.0	77.0
ResNeXt-5062	100.0	77.3
ResNeXt-5112	100.0	77.6
ResNeXt-5162	100.0	77.9
ResNeXt-5212	100.0	78.2
ResNeXt-5262	100.0	78.5
ResNeXt-5312	100.0	78.8
ResNeXt-5362	100.0	79.1
ResNeXt-5412	100.0	79.4
ResNeXt-5462	100.0	79.7
ResNeXt-5512	100.0	80.0
ResNeXt-5562	100.0	80.3
ResNeXt-5612	100.0	80.6
ResNeXt-5662	100.0	80.9
ResNeXt-5712	100.0	81.2
ResNeXt-5762	100.0	81.5
ResNeXt-5812	100.0	81.8
ResNeXt-5862	100.0	82.1
ResNeXt-5912	100.0	82.4
ResNeXt-5962	100.0	82.7
ResNeXt-6012	100.0	83.0
ResNeXt-6062	100.0	83.3
ResNeXt-6112	100.0	83.6
ResNeXt-6162	100.0	83.9
ResNeXt-6212	100.0	84.2
ResNeXt-6262	100.0	84.5
ResNeXt-6312	100.0	84.8
ResNeXt-6362	100.0	85.1
ResNeXt-6412	100.0	85.4
ResNeXt-6462	100.0	85.7
ResNeXt-6512	100.0	86.0
ResNeXt-6562	100.0	86.3
ResNeXt-6612	100.0	86.6
ResNeXt-6662	100.0	86.9
ResNeXt-6712	100.0	87.2
ResNeXt-6762	100.0	87.5
ResNeXt-6812	100.0	87.8
ResNeXt-6862	100.0	88.1
ResNeXt-6912	100.0	88.4
ResNeXt-6962	100.0	88.7
ResNeXt-7012	100.0	89.0
ResNeXt-7062	100.0	89.3
ResNeXt-7112	100.0	89.6
ResNeXt-7162	100.0	89.9
ResNeXt-7212	100.0	90.2
ResNeXt-7262	100.0	90.5
ResNeXt-7312	100.0	90.8
ResNeXt-7362	100.0	91.1
ResNeXt-7412	100.0	91.4
ResNeXt-7462	100.0	91.7
ResNeXt-7512	100.0	92.0
ResNeXt-7562	100.0	92.3
ResNeXt-7612	100.0	92.6
ResNeXt-7662	100.0	92.9
ResNeXt-7712	100.0	93.2
ResNeXt-7762	100.0	93.5
ResNeXt-7812	100.0	93.8
ResNeXt-7862	100.0	94.1
ResNeXt-7912	100.0	94.4
ResNeXt-7962	100.0	94.7
ResNeXt-8012	100.0	95.0
ResNeXt-8062	100.0	95.3
ResNeXt-8112	100.0	95.6
ResNeXt-8162	100.0	95.9
ResNeXt-8212	100.0	96.2
ResNeXt-8262	100.0	96.5
ResNeXt-8312	100.0	96.8
ResNeXt-8362	100.0	97.1
ResNeXt-8412	100.0	97.4
ResNeXt-8462	100.0	97.7
ResNeXt-8512	100.0	98.0
ResNeXt-8562	100.0	98.3
ResNeXt-8612	100.0	98.6
ResNeXt-8662	100.0	98.9
ResNeXt-8712	100.0	99.2
ResNeXt-8762	100.0	99.5
ResNeXt-8812	100.0	99.8
ResNeXt-8862	100.0	100.0
ResNeXt-8912	100.0	100.0
ResNeXt-8962	100.0	100.0
ResNeXt-9012	100.0	100.0
ResNeXt-9062	100.0	100.0
ResNeXt-9112	100.0	100.0
ResNeXt-9162	100.0	100.0
ResNeXt-9212	100.0	100.0
ResNeXt-9262	100.0	100.0
ResNeXt-9312	100.0	100.0
ResNeXt-9362	100.0	100.0
ResNeXt-9412	100.0	100.0
ResNeXt-9462	100.0	100.0
ResNeXt-9512	100.0	100.0
ResNeXt-9562	100.0	100.0
ResNeXt-9612	100.0	100.0
ResNeXt-9662	100.0	100.0
ResNeXt-9712	100.0	100.0
ResNeXt-9762	100.0	100.0
ResNeXt-9812	100.0	100.0
ResNeXt-9862	100.0	100.0
ResNeXt-9912	100.0	100.0
ResNeXt-9962	100.0	100.0
ResNeXt-10012	100.0	100.0
ResNeXt-10062	100.0	100.0
ResNeXt-10112	100.0	100.0
ResNeXt-10162	100.0	100.0
ResNeXt-10212	100.0	100.0
ResNeXt-10262	100.0	100.0
ResNeXt-10312	100.0	100.0
ResNeXt-10362	100.0	100.0
ResNeXt-10412	100.0	100.0
ResNeXt-10462	100.0	100.0
ResNeXt-10512	100.0	100.0
ResNeXt-10562	100.0	100.0
ResNeXt-10612	100.0	100.0
ResNeXt-10662	100.0	100.0
ResNeXt-10712	100.0	100.0
ResNeXt-10762	100.0	100.0
ResNeXt-10812	100.0	100.0
ResNeXt-10862	100.0	100.0
ResNeXt-10912	100.0	100.0
ResNeXt-10962	100.0	100.0
ResNeXt-11012	100.0	100.0
ResNeXt-11062	100.0	100.0
ResNeXt-11112	100.0	100.0
ResNeXt-11162	100.0	100.0
ResNeXt-11212	100.0	100.0
ResNeXt-11262	100.0	100.0
ResNeXt-11312	100.0	100.0
ResNeXt-11362	100.0	100.0
ResNeXt-11412	100.0	100.0
ResNeXt-11462	100.0	100.0
ResNeXt-11512	100.0	100.0
ResNeXt-11562	100.0	100.0
ResNeXt-11612	100.0	100.0
ResNeXt-11662	100.0	100.0
ResNeXt-11712	100.0	100.0
ResNeXt-11762	100.0	100.0
ResNeXt-11812	100.0	100.0
ResNeXt-11862	100.0	100.0
ResNeXt-11912	100.0	100.0
ResNeXt-11962	100.0	100.0
ResNeXt-12012	100.0	100.0
ResNeXt-12062	100.0	100.0
ResNeXt-12112	100.0	100.0
ResNeXt-12162	100.0	100.0
ResNeXt-12212	100.0	100.0
ResNeXt-12262	100.0	100.0
ResNeXt-12312	100.0	100.0
ResNeXt-12362	100.0	100.0
ResNeXt-12412	100.0	100.0
ResNeXt-12462	100.0	100.0
ResNeXt-12512	100.0	100.0
ResNeXt-12562	100.0	100.0
ResNeXt-12612	100.0	100.0
ResNeXt-12662	100.0	100.0
ResNeXt-12712	100.0	100.0
ResNeXt-12762	100.0	100.0
ResNeXt-12812	100.0	100.0
ResNeXt-12862	100.0	100.0
ResNeXt-12912	100.0	100.0
ResNeXt-12962	100.0	100.0
ResNeXt-13012	100.0	100.0
ResNeXt-13062	100.0	100.0
ResNeXt-13112	100.0	100.0
ResNeXt-13162	100.0	100.0

Method	Dim	UCF-101	HMDB-51
ResNeXt-101	3x	90.7	63.8
ResNeXt-101 (64d)	3x	94.5	70.2
CD ² [23]	3x	82.3	-
P3D [24]	3x	82.3	-
Two-stream TD [11]	3x	98.0	80.7
Two-stream CNN [20]	2x	88.0	59.4
TD ² [7]	2x	90.3	63.2
3D Multi-Scale Net [7]	3x	92.0	64.0
TSN [25]	2x	94.2	60.9

森を見る (速読)

ポイントと意外に阻まれる

一 関連研究の最後に従来との比較 差分が書かれる (ごとか多い)

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

Kensho Hara, Hirokazu Kataoka, Yutaka Sato
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
(hara.kensho, kataoka.hirokazu, sato.yutaka@aist.go.jp)

Abstract

The purpose of this study is to determine whether current video datasets have sufficient data for training very deep convolutional neural networks (CNNs) with spatiotemporal three-dimensional (3D) features. Recently, the performance tests of 3D CNNs in the field of action recognition have improved significantly. However, in fact, conventional research has only explored relatively shallow 3D architectures. We examine the architecture of various 3D CNNs from relatively shallow to very deep ones on current video datasets. Based on the results of these experiments, the following conclusions could be obtained: (1) ReNet-18 training resulted in significant overfitting on UCF-101, HMDB-51, and ActivityNet but not for Kinetics. (2) The Kinetics dataset has sufficient data for training of deep 3D CNNs. (3) ReNet-18 is not overfitted on UCF-101 and HMDB-51, respectively. The use of 3D CNNs trained on ImageNet has produced superior performance in action tasks. We believe that using deep 3D CNNs together with Kinetics will retrace the prehistory of 2D CNNs and ImageNet, and stimulate advances in video datasets for video action recognition. We present models used in this study and publicly available.



Figure 1: Recent advances in computer vision for images (top) and video datasets (bottom). The use of very deep 3D CNNs trained on ImageNet produces outstanding progress in image recognition as well as in action tasks. Can the use of 3D CNNs trained on Kinetics generate similar progress in computer vision tasks?

1. Introduction

The use of large-scale datasets is extremely important when using deep convolutional neural networks (CNNs), which have intuitive parameter numbers, and the use of CNNs in the field of computer vision has expanded significantly in recent years. ImageNet [1], which includes more

<https://github.com/kenshohara/3d-kinetics-fork>

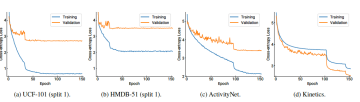


Figure 4: ReNet-18 training and validation losses. The validation losses on UCF-101, HMDB-51, and ActivityNet quickly converged to high values and were clearly higher than their corresponding training losses. The validation losses on Kinetics were slightly higher than the corresponding training losses, significantly different from the other datasets.

trained. The average duration of each video is about 3 seconds. These training splits (CNN training and 30% testing) are provided in this study. The Kinetics dataset provides samples from 200 human action classes with an average of 17 annotated videos per class and 114 action instances per video. Unlike the other datasets, which include videos of untrained videos, which include unaction frames. The total video length is 649 hours, and the total number of action instances is 28,108. This dataset is randomly split into three different subsets: Kinetics-100 for validation and testing. More specifically, 30% is used for training, 25% is used for validation, and 25% is used for testing.

The Kinetics dataset has 600 human action classes, and consists of more than 400 videos for each class. The videos were temporally trained and last around 10 seconds. The total number of the videos is 10,000. The number of training, validation, and testing sets are about 100, 100, and 40,000, respectively.

The video properties of all three datasets are similar. Most videos were extracted from YouTube, except for HMDB-51, which includes videos extracted from Kinetics. The videos include diverse background and camera motion, and some difference among them in the numbers of action classes and instances.

We trained the videos in lengths of 200 pixels without changing their aspect ratios and there used them.

4. Results and discussion

4.1. Analyses of training on each dataset

We began by training ReNet-18 on each dataset. According to previous works [1, 3], 3D CNNs trained on UCF-101, HMDB-51, and ActivityNet do not achieve high accuracy results trained in Kinetics work well. We tried to reproduce such results in this experiment. In this process, we used split 1 of UCF-101 and HMDB-51, which was the training and validation sets of ActivityNet and Kinetics.

is a somewhat larger video dataset, has become available, and in fact has made it possible to accomplish additional tasks such as supervised action classification and detection, but the number of action instances is constant as still limited. More recently, the Kinetics dataset [15] was created with the aim of being positioned as a large video dataset intended that it is roughly equivalent to the position held by ImageNet in relation to image datasets. More than 300k videos have been collected for the Kinetics dataset, which means that the scale of video datasets has begun to approach that of image datasets.

For action recognition, CNNs with spatiotemporal three-dimensional (3D) convolutional kernels (3D CNNs) are recently more effective than CNNs with two-dimensional (2D) kernels [1]. From several years ago [14, 16], 3D CNNs are thought to be possible as effective tool for action recognition. However, even the usage of well-organized networks [12, 13] has failed to overcome the advantage of 2D-based CNNs that combine both spatial and RGB images [12]. The primary reason for this failure has been the relatively small data-scale of video datasets that are available for optimizing the immense number of parameters in 3D CNNs, which are much larger than those of 2D CNNs. In addition, baseline 3D CNNs can only be trained on video datasets, which are much smaller than those of ImageNet. Recently, however, Carreira and Zisserman achieved a significant breakthrough using the Kinetics dataset as well as the ImageNet 2D kernels pretrained on ImageNet to train 3D CNNs [17]. This work has the benefit of a substantial 3D convolution that can be engaged by the Kinetics dataset.

However, can 3D CNNs retrace the successful history of 2D CNNs? More specifically, can the use of 3D CNNs trained on Kinetics produce significant progress in action recognition and other various tasks? (See bottom right of Figure 2.) In order to answer this question, we trained on such large-scale datasets, a large number of algorithms, such as residual learning [17], have been used to improve classification performance by adding residual flow to CNNs, and the use of very deep CNNs trained on ImageNet has facilitated the acquisition of general feature representations. Using such feature representation, in turn, ImageNet has improved the performance of several other tasks including object detection, semantic segmentation, and image captioning (see top left in Figure 1).

In fact, the video datasets available for action recognition have been drastically smaller than compared with image datasets. Representative video datasets, such as UCF-101 [12] and HMDB-51 [17], can be used to provide realistic videos with sizes around 10s. But even though they are still used as standard benchmarks, such datasets are obviously too small to be used for optimizing CNN representations from scratch. In the case couple of years, ActivityNet [13], which

the accuracy of ReNet-18. This result ReNet-200 started to overfit, similar to the results of ReNet-101, which occurred at the depth increased 152, and then the accuracy decreased to the depth of 200. These results indicate that insufficient data for training on ImageNet.

Comparing the results of ReNet-18, ReNet-101, and ReNet-200 are shown in Figure 2. Here, it can be seen that the accuracy of ReNet-200 is slightly higher than that of ReNet-101. In order to provide the above question, the 3D CNN is this study based on results, and then the accuracy of ReNet-200 is slightly higher than that of ReNet-101. This result is similar to the results of ReNet-101, which occurred at the depth increased 152, and then the accuracy decreased to the depth of 200. These results indicate that insufficient data for training on ImageNet.

Comparing the results of ReNet-18, ReNet-101, and ReNet-200 are shown in Figure 2. Here, it can be seen that the accuracy of ReNet-200 is slightly higher than that of ReNet-101. In order to provide the above question, the 3D CNN is this study based on results, and then the accuracy of ReNet-200 is slightly higher than that of ReNet-101. This result is similar to the results of ReNet-101, which occurred at the depth increased 152, and then the accuracy decreased to the depth of 200. These results indicate that insufficient data for training on ImageNet.

Comparing the results of ReNet-18, ReNet-101, and ReNet-200 are shown in Figure 2. Here, it can be seen that the accuracy of ReNet-200 is slightly higher than that of ReNet-101. In order to provide the above question, the 3D CNN is this study based on results, and then the accuracy of ReNet-200 is slightly higher than that of ReNet-101. This result is similar to the results of ReNet-101, which occurred at the depth increased 152, and then the accuracy decreased to the depth of 200. These results indicate that insufficient data for training on ImageNet.

Comparing the results of ReNet-18, ReNet-101, and ReNet-200 are shown in Figure 2. Here, it can be seen that the accuracy of ReNet-200 is slightly higher than that of ReNet-101. In order to provide the above question, the 3D CNN is this study based on results, and then the accuracy of ReNet-200 is slightly higher than that of ReNet-101. This result is similar to the results of ReNet-101, which occurred at the depth increased 152, and then the accuracy decreased to the depth of 200. These results indicate that insufficient data for training on ImageNet.

Comparing the results of ReNet-18, ReNet-101, and ReNet-200 are shown in Figure 2. Here, it can be seen that the accuracy of ReNet-200 is slightly higher than that of ReNet-101. In order to provide the above question, the 3D CNN is this study based on results, and then the accuracy of ReNet-200 is slightly higher than that of ReNet-101. This result is similar to the results of ReNet-101, which occurred at the depth increased 152, and then the accuracy decreased to the depth of 200. These results indicate that insufficient data for training on ImageNet.

including 250,000 action instances. ActivityNet also provides some additional tasks, such as unsupervised classification and detection, but the number of action instances is still as the order of tens of thousands. This year (2017), an effort to create a successful pretrained model, Kay et al. released the Kinetics dataset [15]. The Kinetics dataset includes more than 300,000 training videos covering 400 classes, in order to determine whether it can train deeper 3D CNNs.

Other large datasets such as Sports-1M [13] and YouTube-8M [1] have been proposed. Although these datasets are larger than Kinetics, their annotations are slightly noisy and only video-level labels have been assigned. In other words, they include frames that do not belong to target actions. Such noise and the presence of unrelated frames have the potential to prevent these models from proceeding good training. In addition, with the size of 10s of 10 TB, these scales are simply too large to allow them to be utilized easily. Because of these issues, as well as the fact that the Kinetics dataset is more suitable for training deep 3D CNNs than scratch, we used the Kinetics dataset in this study.

As for the fine-tuning. The results of these experiments (Section 4 for details) show the Kinetics dataset is similar to the ImageNet 2D kernels pretrained on ImageNet, as the training accomplished by 2D ReNet-18/ImageNet, as shown in Figure 2. Based on these results, we decided to use the ImageNet 2D kernels pretrained on ImageNet to train 3D CNNs. In this study, Simonyan et al. proposed a method to use RGB and stacked optical flow frames as separate and merge information, respectively [18], and showed that combining the two-nets has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-nets CNNs have been proposed to improve action recognition performance [1, 3, 17, 22, 23].

Until the aforementioned approaches, we focused on CNNs with 2D convolutional kernels, which have recently been incorporated into 3D CNNs through the use of large-scale video datasets. These 3D CNNs are intuitively effective because such 3D convolution can be used to directly extract spatiotemporal features from raw videos. For example, Si et al. proposed applying 3D convolutions to extract spatiotemporal features from videos, while Tran et al. trained 3D CNNs, which they referred to as C3D, using the Sports-1M dataset [13]. Since that study, C3D has been used as a

2.2. Action Recognition Approaches

One of the popular approaches to CNN-based action recognition is to use RGB and stacked optical flow frames as separate and merge information, respectively [18], and showed that combining the two-nets has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-nets CNNs have been proposed to improve action recognition performance [1, 3, 17, 22, 23].

Until the aforementioned approaches, we focused on CNNs with 2D convolutional kernels, which have recently been incorporated into 3D CNNs through the use of large-scale video datasets. These 3D CNNs are intuitively effective because such 3D convolution can be used to directly extract spatiotemporal features from raw videos. For example, Si et al. proposed applying 3D convolutions to extract spatiotemporal features from videos, while Tran et al. trained 3D CNNs, which they referred to as C3D, using the Sports-1M dataset [13]. Since that study, C3D has been used as a

2.2.1. Video Datasets

One of the popular approaches to CNN-based action recognition is to use RGB and stacked optical flow frames as separate and merge information, respectively [18], and showed that combining the two-nets has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-nets CNNs have been proposed to improve action recognition performance [1, 3, 17, 22, 23].

Until the aforementioned approaches, we focused on CNNs with 2D convolutional kernels, which have recently been incorporated into 3D CNNs through the use of large-scale video datasets. These 3D CNNs are intuitively effective because such 3D convolution can be used to directly extract spatiotemporal features from raw videos. For example, Si et al. proposed applying 3D convolutions to extract spatiotemporal features from videos, while Tran et al. trained 3D CNNs, which they referred to as C3D, using the Sports-1M dataset [13]. Since that study, C3D has been used as a

2.2.2. Action Recognition Approaches

One of the popular approaches to CNN-based action recognition is to use RGB and stacked optical flow frames as separate and merge information, respectively [18], and showed that combining the two-nets has the ability to improve action recognition accuracy. Since that study, numerous methods based on the two-nets CNNs have been proposed to improve action recognition performance [1, 3, 17, 22, 23].

Until the aforementioned approaches, we focused on CNNs with 2D convolutional kernels, which have recently been incorporated into 3D CNNs through the use of large-scale video datasets. These 3D CNNs are intuitively effective because such 3D convolution can be used to directly extract spatiotemporal features from raw videos. For example, Si et al. proposed applying 3D convolutions to extract spatiotemporal features from videos, while Tran et al. trained 3D CNNs, which they referred to as C3D, using the Sports-1M dataset [13]. Since that study, C3D has been used as a

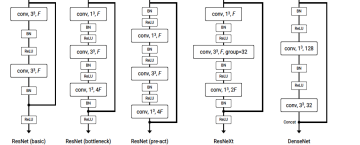
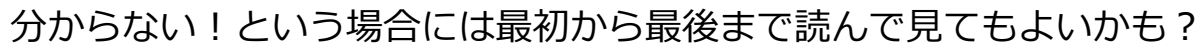


Figure 3: Block of each architecture. We represent conv_1, conv_2, ..., conv_5 as the kernel size, and the number of feature maps of the convolutional filter as $s \times s \times c$, respectively, and group the number of groups of group-convolutions, which divide the feature maps into small groups. BN refers to batch normalization [1]. Shallow connections of the architectures are annotated except for those of DenseNet, which omit concatenation.

Table 1. Network Architectures. Each convolutional layer is followed by batch normalization [1] and a ReLU [10]. Spatio-temporal down-sampling is performed by conv_1, conv_2, ..., conv_5, except for DenseNet. P is the number of feature channels. The number of feature channels is shown in Figure 1, and V is the number of feature channels. DenseNet down-samples inputs using the transition layer, which consists of $3 \times 3 \times 3$ convolutional layer and $2 \times 2 \times 2$ average pooling layer with a stride of one, after conv_2, conv_3, ..., and conv_5. F and N are the number of input feature channels of each block in each layer, and V is the same as that of the other networks. A $3 \times 3 \times 3$ non-grouping layer (conv_1) is also located before conv_2. c of all networks for down-sampling. In addition, conv_1 is applied to each input with a spatial stride of 2. c of the fully-connected layer is the number of classes.

Model	conv1	conv2	conv3	conv4	conv5	conv6	conv7	conv8	conv9	conv10	conv11	conv12	conv13	conv14	conv15	conv16	conv17	conv18	conv19	conv20	conv21	conv22	conv23	conv24	conv25	conv26	conv27	conv28	conv29	conv30	conv31	conv32	conv33	conv34	conv35	conv36	conv37	conv38	conv39	conv40	conv41	conv42	conv43	conv44	conv45	conv46	conv47	conv48	conv49	conv50	conv51	conv52	conv53	conv54	conv55	conv56	conv57	conv58	conv59	conv60	conv61	conv62	conv63	conv64	conv65	conv66	conv67	conv68	conv69	conv70	conv71	conv72	conv73	conv74	conv75	conv76	conv77	conv78	conv79	conv80	conv81	conv82	conv83	conv84	conv85	conv86	conv87	conv88	conv89	conv90	conv91	conv92	conv93	conv94	conv95	conv96	conv97	conv98	conv99	conv100	conv101	conv102	conv103	conv104	conv105	conv106	conv107	conv108	conv109	conv110	conv111	conv112	conv113	conv114	conv115	conv116	conv117	conv118	conv119	conv120	conv121	conv122	conv123	conv124	conv125	conv126	conv127	conv128	conv129	conv130	conv131	conv132	conv133	conv134	conv135	conv136	conv137	conv138	conv139	conv140	conv141	conv142	conv143	conv144	conv145	conv146	conv147	conv148	conv149	conv150	conv151	conv152	conv153	conv154	conv155	conv156	conv157	conv158	conv159	conv160	conv161	conv162	conv163	conv164	conv165	conv166	conv167	conv168	conv169	conv170	conv171	conv172	conv173	conv174	conv175	conv176	conv177	conv178	conv179	conv180	conv181	conv182	conv183	conv184	conv185	conv186	conv187	conv188	conv189	conv190	conv191	conv192	conv193	conv194	conv195	conv196	conv197	conv198	conv199	conv200	conv201	conv202	conv203	conv204	conv205	conv206	conv207	conv208	conv209	conv210	conv211	conv212	conv213	conv214	conv215	conv216	conv217	conv218	conv219	conv220	conv221	conv222	conv223	conv224	conv225	conv226	conv227	conv228	conv229	conv230	conv231	conv232	conv233	conv234	conv235	conv236	conv237	conv238	conv239	conv240	conv241	conv242	conv243	conv244	conv245	conv246	conv247	conv248	conv249	conv250	conv251	conv252	conv253	conv254	conv255	conv256	conv257	conv258	conv259	conv260	conv261	conv262	conv263	conv264	conv265	conv266	conv267	conv268	conv269	conv270	conv271	conv272	conv273	conv274	conv275	conv276	conv277	conv278	conv279	conv280	conv281	conv282	conv283	conv284	conv285	conv286	conv287	conv288	conv289	conv290	conv291	conv292	conv293	conv294	conv295	conv296	conv297	conv298	conv299	conv300	conv301	conv302	conv303	conv304	conv305	conv306	conv307	conv308	conv309	conv310	conv311	conv312	conv313	conv314	conv315	conv316	conv317	conv318	conv319	conv320	conv321	conv322	conv323	conv324	conv325	conv326	conv327	conv328	conv329	conv330	conv331	conv332	conv333	conv334	conv335	conv336	conv337	conv338	conv339	conv340	conv341	conv342	conv343	conv344	conv345	conv346	conv347	conv348	conv349	conv350	conv351	conv352	conv353	conv354	conv355	conv356	conv357	conv358	conv359	conv360	conv361	conv362	conv363	conv364	conv365	conv366	conv367	conv368	conv369	conv370	conv371	conv372	conv373	conv374	conv375	conv376	conv377	conv378	conv379	conv380	conv381	conv382	conv383	conv384	conv385	conv386	conv387	conv388	conv389	conv390	conv391	conv392	conv393	conv394	conv395	conv396	conv397	conv398	conv399	conv400	conv401	conv402	conv403	conv404	conv405	conv406	conv407	conv408	conv409	conv410	conv411	conv412	conv413	conv414	conv415	conv416	conv417	conv418	conv419	conv420	conv421	conv422	conv423	conv424	conv425	conv426	conv427	conv428	conv429	conv430	conv431	conv432	conv433	conv434	conv435	conv436	conv437	conv438	conv439	conv440	conv441	conv442	conv443	conv444	conv445	conv446	conv447	conv448	conv449	conv450	conv451	conv452	conv453	conv454	conv455	conv456	conv457	conv458	conv459	conv460	conv461	conv462	conv463	conv464	conv465	conv466	conv467	conv468	conv469	conv470	conv471	conv472	conv473	conv474	conv475	conv476	conv477	conv478	conv479	conv480	conv481	conv482	conv483	conv484	conv485	conv486	conv487	conv488	conv489	conv490	conv491	conv492	conv493	conv494	conv495	conv496	conv497	conv498	conv499	conv500	conv501	conv502	conv503	conv504	conv505	conv506	conv507	conv508	conv509	conv510	conv511	conv512	conv513	conv514	conv515	conv516	conv517	conv518	conv519	conv520	conv521	conv522	conv523	conv524	conv525	conv526	conv527	conv528	conv529	conv530	conv531	conv532	conv533	conv534	conv535	conv536	conv537	conv538	conv539	conv540	conv541	conv542	conv543	conv544	conv545	conv546	conv547	conv548	conv549	conv550	conv551	conv552	conv553	conv554	conv555	conv556	conv557	conv558	conv559	conv560	conv561	conv562	conv563	conv564	conv565	conv566	conv567	conv568	conv569	conv570	conv571	conv572	conv573	conv574	conv575	conv576	conv577	conv578	conv579	conv580	conv581	conv582	conv583	conv584	conv585	conv586	conv587	conv588	conv589	conv590	conv591	conv592	conv593	conv594	conv595	conv596	conv597	conv598	conv599	conv600	conv601	conv602	conv603	conv604	conv605	conv606	conv607	conv608	conv609	conv610	conv611	conv612	conv613	conv614	conv615	conv616	conv617	conv618	conv619	conv620	conv621	conv622	conv623	conv624	conv625	conv626	conv627	conv628	conv629	conv630	conv631	conv632	conv633	conv634	conv635	conv636	conv637	conv638	conv639	conv640	conv641	conv642	conv643	conv644	conv645	conv646	conv647	conv648	conv649	conv650	conv651	conv652	conv653	conv654	conv655	conv656	conv657	conv658	conv659	conv660	conv661	conv662	conv663	conv664	conv665	conv666	conv667	conv668	conv669	conv670	conv671	conv672	conv673	conv674	conv675	conv676	conv677	conv678	conv679	conv680	conv681	conv682	conv683	conv684	conv685	conv686	conv687	conv688	conv689	conv690	conv691	conv692	conv693	conv694	conv695	conv696	conv697	conv698	conv699	conv700	conv701	conv702	conv703	conv704	conv705	conv706	conv707	conv708	conv709	conv710	conv711	conv712	conv713	conv714	conv715	conv716	conv717	conv718	conv719	conv720	conv721	conv722	conv723	conv724	conv725	conv726	conv727	conv728	conv729	conv730	conv731	conv732	conv733	conv734	conv735	conv736	conv737	conv738	conv739	conv740	conv741	conv742	conv743	conv744	conv745	conv746	conv747	conv748	conv749	conv750	conv751	conv752	conv753	conv754	conv755	conv756	conv757	conv758	conv759	conv760	conv761	conv762	conv763	conv764	conv765	conv766	conv767	conv768	conv769	conv770	conv771	conv772	conv773	conv774	conv775	conv776	conv777	conv778	conv779	conv780	conv781	conv782	conv783	conv784	conv785	conv786	conv787	conv788	conv789	conv790	conv791	conv792	conv793	conv794	conv795	conv796	conv797	conv798	conv799	conv800	conv801	conv802	conv803	conv804	conv805	conv806	conv807	conv808	conv809	conv810	conv811	conv812	conv813	conv814	conv815	conv816	conv817	conv818	conv819	conv820	conv821	conv822	conv823	conv824	conv825	conv826	conv827	conv828	conv829	conv830	conv831	conv832	conv833	conv834	conv835	conv836	conv837	conv838	conv839	conv840	conv841	conv842	conv843	conv844	conv845	conv846	conv847	conv848	conv849	conv850	conv851	conv852	conv853	conv854	conv855	conv856	conv857	conv858	conv859	conv860	conv861	conv862	conv863	conv864	conv865	conv866	conv867	conv868	conv869	conv870	conv871	conv872	conv873	conv874	conv875	conv876	conv877	conv878	conv879	conv880	conv881	conv882	conv883	conv884	conv885	conv886	conv887	conv888	conv889	conv890	conv891	conv892	conv893	conv894	conv895	conv896	conv897	conv898	conv899	conv900	conv901	conv902	conv903	conv904	conv905	conv906	conv907	conv908	conv909	conv910	conv911	conv912	conv913	conv914	conv915	conv916	conv917	conv918	conv919	conv920	conv921	conv922	conv923	conv924	conv925	conv926	conv927	conv928	conv929	conv930	conv931	conv932	conv933	conv934	conv935	conv936	conv937	conv938	conv939	conv940	conv941	conv942	conv943	conv944	conv945	conv946	conv947	conv948	conv949	conv950	conv951	conv952	conv953	conv954	conv955	conv956	conv957	conv958	conv959	conv960	conv961	conv962	conv963	conv964	conv965	conv966	conv967	conv968	conv969	conv970	conv971	conv972	conv973	conv974	conv975	conv976	conv977	conv978	conv979	conv980	conv981	conv982	conv983	conv984	conv985	conv986	conv987	conv988	conv989	conv990	conv991	conv992	conv993	conv994	conv995	conv996	conv997	conv998	conv999	conv1000	conv1001	conv1002	conv1003	conv1004	conv1005	conv1006	conv1007	conv1008	conv1009	conv1010	conv1011	conv1012	conv1013	conv1014	conv1015	conv1016	conv1017	conv1018	conv1019	conv1020	conv1021	conv1022	conv1023	conv1024	conv1025	conv1026	conv1027	conv1028	conv1029	conv1030	conv1031	conv1032	conv1033	conv1034	conv1035	conv1036	conv1037	conv1038	conv1039	conv1040	conv1041	conv1042	conv1043	conv1044	conv1045	conv1046	conv1047	conv1048	conv1049	conv1050	conv1051	conv1052	conv1053	conv1054	conv1055	conv1056	conv1057	conv1058	conv1059	conv1060	conv1061	conv1062	conv1063	conv1064	conv1065	conv1066	conv1067	conv1068	conv1069	conv1070	conv1071	conv1072	conv1073	conv1074	conv1075	conv1076	conv1077	conv1078	conv1079	conv1080	conv1081	conv1082	conv1083	conv1084	conv1085	conv1086	conv1087	conv1088	conv1089	conv1090	conv1091	conv1092	conv1093	conv1094	conv1095	conv1096	conv1097	conv1098	conv1099	conv1100	conv1101	conv1102	conv1103	conv1104	conv1105	conv1106	conv1107	conv1108	conv1109	conv1110	conv1111	conv1112	conv1113	conv1114	conv1115	conv1116	conv1117	conv1118	conv1119	conv1120	conv1121	conv1122	conv1123	conv1124	conv1125	conv1126	conv1127	conv1128	conv1129	conv1130	conv1131	conv1132	conv1133	conv1134	conv1135	conv1136	conv1137	conv1138	conv1139	conv1140	conv1141	conv1142	conv1143	conv1144	conv1145	conv1146	conv1147	conv1148	conv1149	conv1150	conv1151	conv1152	conv1153	conv1154	conv1155	conv1156	conv1157	conv1158	conv1159	conv1160	conv1161	conv1162	conv1163	conv1164	conv1165	conv1166	conv1167	conv1168	conv1169	conv1170	conv1171	conv1172	conv1173	conv1174	conv1175	conv1176	conv1177	conv1178	conv1179	conv1180	conv1181	conv1182	conv1183	conv1184	conv1185	conv1186	conv1187	conv1188	conv1189	conv1190	conv1191	conv1192	conv1193	conv1194	conv1195	conv1196	conv1197	conv1198	conv1199	conv1200	conv1201	conv1202	conv1203	conv1204	conv1205	conv1206	conv1207	conv1208	conv1209	conv1210	conv1211	conv1212	conv1213	conv1214	conv1215	conv1216	conv1217	conv1218	conv1219	conv1220	conv1221	conv1222	conv1223	conv1224	conv1225	conv1226	conv1227	conv1228	conv1229	conv1230	conv1231	conv1232	conv1233	conv1234	conv1235	conv1236	conv1237	conv1238	conv1239	conv1240	conv1241	conv1242	conv1243	conv1244	conv1245	conv1246	conv1247	conv1248	conv1249	conv1250	conv1251	conv1252	conv1253	conv1254	conv1255	conv1256	conv1257	conv1258	conv1259	conv1260	conv1261	conv1262	conv1263	conv1264	conv1265	conv1266	conv1267	conv1268	conv1269	conv1270	conv1271	conv1272	conv1273	conv1274	conv1275	conv1276	conv1277	conv1278	conv1279	conv1280	conv1281	conv1282	conv1283	conv1284	conv1285	conv1286	conv1287	conv1288	conv1289	conv1290	conv1291	conv1292	conv1293	conv1294	conv1295	conv1296	conv1297	conv1298	conv1299	conv1300	conv1301	conv1302	conv1303	conv1304	conv1305	conv1306	conv1307	conv1308	conv1309	conv1310	conv1311	conv1312	conv1313	conv1314	conv1315	conv1316	conv1317	conv1318	conv1319	conv1320	conv1321	conv1322	conv1323	conv1324	conv1325	conv1326	conv1327	conv1328	conv1329	conv1330	conv1331	conv1332	conv1333	conv1334	conv1335	conv1336	conv1337	conv1338	conv1339	conv1340	conv1341	conv1342	conv1343	conv1344	conv1345	conv1346	conv1347	conv1348
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	---------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

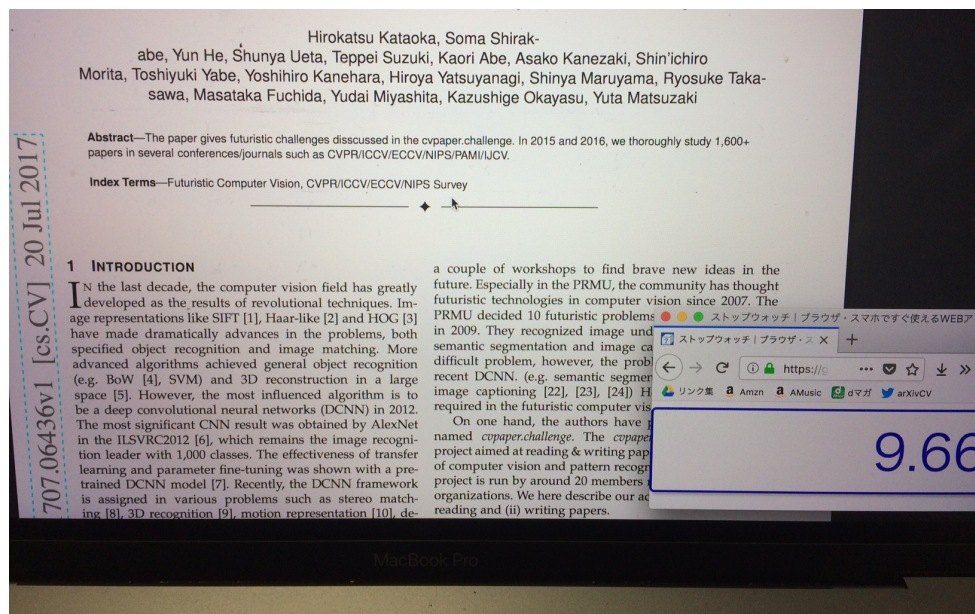
- 基本的には速読と、部分を詳細に読み込んでいく
- ここから先は何をしたいかに依存する



気をつけていること

タイムトライアル

- 時間を気にして、締め切りある読みにする
 - 実際ストップウォッチにおいて論文読んでます！
- 目安時間
 - 速読（15～30分）
 - 精読（1時間～理解できるまで）



気をつけていること

まとめを作成しよう

- （自分だけでなく他の人が）素早く研究の肝をつかむ
- まとめることで記憶の定着を早くする

【1】  Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, Antonio Torralba, "SUN Database: Large-scale Scene Recognition from Abbey to Zoo", in CVPR2010.

Keywords: Dataset, Scene Categorization, Benchmark, Recognition

概要	データセットの概要
コンピュータビジョンにおいてシーン認識のデータベースである Scene UNderstanding (SUN) databaseを提案。シーン認識の裾野を広げた。	シーン認識に関する397クラス、130,519枚の画像が含まれる。画像例は次ページ。比較した特徴量は、HOG, denseSIFT, self-similarity (ssim), LBP, GIST, textonなど。
新規性・差分	結果
それまでの物体認識のデータセットでは数百クラスの識別クラスが用意されていたが、シーン認識では15種類程度しか含まれていなかった。SUN databaseでは、それまでのデータセットをさらに拡大させ、397クラスのシーンを含む、大規模なデータセットである。	次ページの図の通り。全ての特徴量を統合するのが最も精度が高いことが判明した(38.0%)。次いでHOG2x2 (27.2%), geometry texton hist (23.5%), ssim (22.5%), dense SIFT (21.5%)であった。
Links	
論文ページ: http://cs.brown.edu/~hays/papers/sun.pdf	HOG https://hal.archives-ouvertes.fr/inria-00548512/document GIST http://cyci.mit.edu/scene_understanding.html SSIM http://www.researchgate.net/profile/Eli_Shechtman/publication/221362526_Matching_Local_Self-Similarities_across_Images_and_Videos/links/02e7e520897af25746000000.pdf DenseSIFT http://www.vision.caltech.edu/Image_Datasets/Caltech101/cvpr06b_lana.pdf LBP http://www.outex.oulu.fi/publications/pami_02_opm.pdf Sparse SIFT http://www.robots.ox.ac.uk/~vgg/publications/papers/sivic04b.pdf Texton http://www.ics.uci.edu/~fowlkes/papers/mftm-iccv01.pdf
プロジェクトページ: http://vision.princeton.edu/projects/2010/SUN/	

まとめの例：概要、新規性、手法、結果、リンク等があるのが望ましい

グループサーベイで意識すること

集団の力をうまく利用する

- 「個人」でのサーベイ
 - 上記の通り
- 「グループ」でのサーベイ
 - 分割と統合をうまく活用して知識を共有
 - 分担して資料を作成してお互いに参照
 - 同じ論文を読んで補完

最近の重要論文は本数が多いので協力して知識を獲得する

ITツールを活用してディスカッション

一人で読むよりも、みんなで協力して進めるのが現代流

- クラウド (Google/Dropbox) , チャット (Slack/Skype/Line)
- モチベーション, 集合知, 作業分割と統合, ディスカッション等グループで活動するメリットは多い

運れはせながらReadingListのURLをCVFのものに置換しました。
今年のCVFのページにはarXivのリンクもついていますね。年次の投稿数遷移が取れるかと思いましたが、ICCV2015にはリンクなかったです。残念。

ワークショップの動向:
若いワークショップが多い?印象です
> ざっくり数えたところ今回が4回目以下のWSが16件/44件
面白そうなワークショップ
> COMPUTER VISION PROBLEMS IN PLANT PHENOTYPING (CVPPP)
> <https://www.plant-phenotyping.org/CVPPP2017>
> 日本は農業に力を入れている印象であり、国内学会(View, SSIIなど)だと農業分野の外観検査関連の研究があるので、このWSは今後ねらい目ではないでしょうか

(edited)
葉っぱセグメンテーション&カウントのチャレンジですね。
かなり成熟した感じのある画像認識技術を、アプリケーション寄りにすそ野を広げたいという意図が働いているのでしょうか？
非CV屋さんを取り込みたい？

GANチュートリアル (スライド含む) link: <https://sites.google.com/view/iccv-2017-gans/schedule>

hirokatsu.kataoka 2:13 PM

1. Mask-RCNN

は前述の通りResNet, Faster RCNNで有名な Kaimingさんの著書です。物体 検出とセマンティックセグメンテーションを同時に解いた方が良い、という知見に基づいています。これ以降、Kaimingさんの成果の速度がよくなっているのは、実はこの論文が (検出やセグメンテーションの) 完成版で、彼はずでに違うことにフォーカスして 研究しているのでは？と見ています。

2. Kd-network

ICCV 2017 速報の資料作成

	Erhan, Vincent Vanhoucke, Andrew Rabinovich		Anton van den Hengel, Chris Russell, Anthony Dick, John Bastian, Daniel Pooley, Lachlan Fleming, Lourdes Agapito
1505	Understanding Image Representations by Measuring Their Equivariance and Equivalence Karel Lenc, Andrea Vedaldi	1505	SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite
			Shuran Song, Samuel P. Lichtenberg, Jianxiong Xiao
1505	Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images Anh Nguyen, Jason Yosinski, Jeff Clune	1505	Small-Variance Nonparametric Clustering on the Hypersphere Julian Straub, Trevor Campbell, Jonathan P. How, John W. Fisher III

Monday June 8, 10:10am-12:30pm


Poster Session	
1505	Going Deeper With Convolutions Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich
1506	Propagated Image Filtering Jen-Hao Rick Chang, Yu-Chiang Frank Wang
1506	Web Scale Photo Hash Clustering on a Single Machine Yunchao Gong, Marcin Pawlowski, Fei Yang, Louis Brandy, Lubomir Bourdev, Rob Fergus
1506	Expanding Object Detector's Horizon: Incremental Learning Framework for Object Detection in Videos Alina Kuznetsova, Sung Ju Hwang, Bodo Rosenhahn, Leonid Sigal
1506	Supervised Discrete Hashing Fumin Shen, Chunhua Shen, Wei Liu, Heng Tao Shen
1505	What do 15,000 Object Categories Tell Us About Classifying and Localizing Actions? Mihir Jain, Jan C. van Gemert, Cees G. M. Snoek
1508	Landmarks-Based Kernelized Subspace Alignment for Unsupervised Domain Adaptation Rahaf Aljundi, Rémi Emonet, Damien Muzet, Marc Sebban

CVPR2015 完全読破


スライドを共有

資料をみんなで作り上げていく

- Ver.を上げていくごとに自分と他人の知見を混ぜていく
- 間接的に読んで、議論を重ねるうちに自分にも知識が定着


 10 PM
簡易版テンプレートでspotlightとoralを13本まとめました。そして、今日現時点の10月まとめ資料をppt資料に挿入・変換しました。現在ICCV2017論文まとめは合計75本です。どうぞ、よろしくお願いいたします。こういった資料を入れてV6からV7をアップロードします。


 **hirokatsu.kataoka** 10:27 PM
uploaded and commented on this file ▼

 **iccv17_update.pptx**
90 kB PowerPoint Presentation

“ 各自のアップデート分のみを更新する資料も準備しました！こちらの該当部分に書き込んで送ってもらえたら更新部分が丸わかりであります。

October 28th, 2017

 1 PM
uploaded and commented on this file ▼

 **iccv17_update_qiu.pptx**
16 MB PowerPoint Presentation

“ 気づきはなかなか書きづらいです。ほぼ論文まとめなので、すみません。

CVPR/ICCV 2017 速報の資料作成

可視化，競争

可視化すると意識して読むようになる（？）

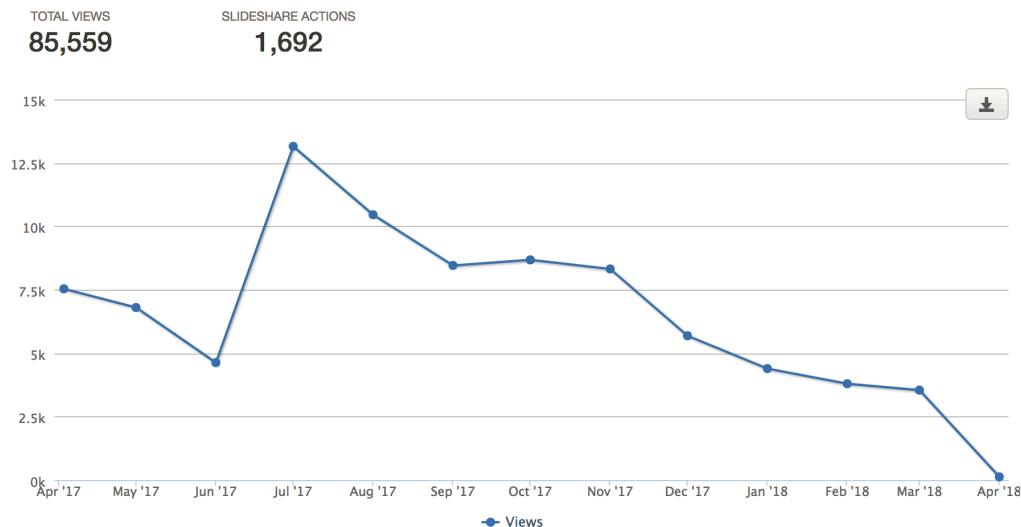
- 週間，月間ランクなど
- 機関総合 (2018/3/1～2018/3/31) 1位：163本, 2位：95本, 3位：87本
- 個人総合 (2018/3/1～2018/3/31) 1位：80本, 2位：68本, 3位：45本

資料公開のススメ

人目に触れて叩き上げる, プレゼンスを高める

- 評価を見て資料の出来栄えを判断（面白いかどうかくらいはわかる）
- 学会などでリアルにあうと反応をもらえる

論文が通った時だけ宣伝, ではなく
普段やっていることでプレゼンスを上げる



Top content

Name	Views
CVPR 2017 速報	22,526
ICCV 2017 速報	9,486
コンピュータビジョンの今を映す-CVPR 2017 速報より- (夏のトップカンファレンス論文読み会)	4,176
【2017.03】 cvpaper.challenge2017	2,649
CVPR 2016 まとめ v1	2,315

サーベイ法まとめ

サーベイの方法論について、個人/グループという単位で説明

- 速読と精読を組み合わせた個人サーベイ
- 組織的サーベイで早く/確実に知識を作り上げていく

もっと大事なこと

サーベイ（に限らず研究）は楽しんでやるもの！こんな楽しいことができるようになった，分からなかったことを知識として明らかにした，というゲーム

MIRU若手Pにご参加ください！

https://sites.google.com/view/miru2018sapporo/wakate_top

今回のテーマは「異分野サーベイ」

- 3/18(日): 募集開始
- 4/23(月): 募集締切
- 4/27(金): グループ分けの発表, 企画1: 異分野サーベイ企画開始.
- 5/17(木)-18(金): サーベイに関するチュートリアル講演 (PRMU研究会@岐阜大学) .
 - スケジュールの詳細はPRMU のウェブサイトに掲載予定です. また, 本講演の様子や資料は企画内容のページにアップロード予定です.



ミルクマ