様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２７　日現在

機関番号：　１４６０３
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１６０２４

研究課題名（和文）Fast, effective and robust person re-identification for large-scale real applications

研究課題名（英文）Fast, effective and robust person re-identification for large-scale real applications

研究代表者

伍　洋（Wu, Yang）

奈良先端科学技術大学院大学・研究推進機構・特任助教

研究者番号：３０７５０５５９

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：知能ビデオ監視を現実的になるように、人物照合モデルをより効果的に、よりスケーラブルに、より迅速に、より堅牢にするソリューションを検討した。実効性を高めるために、辞書学習、メトリック学習、局部性、およびエンドツーエンドの深層学習を探求した。効率とスケーラビリティのために、大量のクラスを扱うときでも、分類を高速に検索できる階層型クラス間構造学習モデルを提案した。堅牢性のために、1つの転移学習モデルと2つの能動学習モデルを設計し、最小限の監督でデータセット間のモデル転送とトレーニングを可能にした。10件の論文が発表され、提案されたすべてのモデルが広範な実験と比較によって評価されている。

研究成果の概要（英文）：We investigated solutions for making person re-identification models more effective, more scalable, faster and more robust, toward being applicable to real world intelligent video surveillance scenarios. For enhancing the effectiveness, we have explored dictionary learning, metric learning, locality, and end-to-end deep learning. For efficiency and scalability, we proposed a hierarchical interclass structure learning model which allows fast search for classification even when handling large amounts of classes. For robustness, we designed one transfer learning model and two active learning models which allows across-dataset model transferring and training with minimum supervision. 10 papers have been published and all the proposed models have been evaluated with extensive experiments and comparisons.

研究分野：　コンピュータビジョン

キーワード：　Person Re-identification　Pattern Recognition　Computer Vision　Machine Learning　Deep Learning　Transfer Learning　Active Learning

１．研究開始当初の背景

Video surveillance has been attracting more and more attention from academic institutions, industrial enterprises, governments, and also the public. Deploying more cameras is just the beginning for improving our safety and security, as it makes it harder and harder to get enough labors for either online monitoring to stop a crime/damage beforehand or offline footage checking to find out the criminals/evidences afterwards. Pursuing fast and reliable intelligent analysis using computers has become unavoidable.

Besides building the infrastructure, intelligent video surveillance industry has mainly developed single camera applications such as video compressing, enhancing, irregularity detection, and extracting some statistics and regions of interest (ROI). However, the more interesting applications with camera networks, which may provide various valuable personalized services, such as finding specific person (criminal, lost child, Alzheimer's patient, lost after disaster, etc.), are still unreachable. The fundamental and critical problem is person re-identification, which means identifying a person again when he/she reappears in the view of a camera after disappearing from the view of the same or another camera in a camera network. In some sense, re-identification has become a bottleneck of the development of intelligent video surveillance industry.

As a research topic, person re-identification has a relatively shorter history (no more than 10 years) than many others. Along with several European countries, Japan has pioneered the exploring of this challenging but attractive topic, and now it gets more and more interests from all over the world. Yoshihisa Ijiri et al. has written a good survey on person re-identification in 2011, and the latest progresses can be found in the first book about this topic (Shaogang Gong et al., "Person Re-Identification", 2014) after the success of the 1st workshop on it.

Besides the continuous efforts and progresses in improving the performance on a few small-scale public datasets which were released years ago, there are clear trends that people get more and more interested in larger and more real-world applications oriented datasets collected with many (not just 2 or a few) cameras. However, besides releasing several large-scale datasets (including the one built by ourselves), as far as we are aware there are no effective efforts in solving this challenging problem, especially when many realistic issues such as efficiency, scalability, robustness and adaption ability need to be concerned.

２．研究の目的

Unlike existing research activities on person re-identification which usually just focus on increasing the still low accuracy of person re-identification on several carefully constructed small-scale closed-world benchmark datasets, we directly target at solutions for large-scale open-world applications, which are expected to be not only effective, but also efficient, scalable and adaptive.

To advance the research and accelerate the process of transferring to real-world applications, we plan to achieve the following two objectives in this study, based on our research experiences and achievements.

(1). Proposing a fast, scalable and effective model for large-scale person re-identification

This stage focuses on efficiency and scalability while pursuing effectiveness.

(2). Making the model adaptive and robust to changing environments

This stage focuses on the generalization abilities of the models, for being transferrable across datasets and applicable to the scenarios where labeled training data is difficult to obtain.

３．研究の方法

(1). Proposing a fast, scalable and effective model for large-scale person re-identification

Collaboratively represent the test data using all the concerned samples from different classes (including the class which the test data belongs to) can lead to induced competition among all the collaborated samples, and therefore make the representation coefficients discriminative for effective classification. This is why collaborative representation has shown very impressive classification performance in our former studies.

We further enhance and extended the model by exploring the following five aspects.

① **Dictionary learning**. It can enhance the discrimination ability of the model by replacing the original samples (used for collaborative representation) by learned dictionary items.

② **Metric Learning**. It can help finding a discriminative metric space that makes the samples easily separable.

③ **Locality**. Locality enables more detailed modeling of the complex data by splitting the whole space into small local regions and then modeling the structure of data in such local areas. It can not only increase the capacity and power of the model, but also allows potentially faster search in local regions.

④ **End-to-end deep learning**. We proposed a temporally-enhanced convolutional network structure for end-to-end video-based person re-identification, which showed the state-of-the-art performance while being computationally efficient.

⑤ **Hierarchical structure**. For speeding up and scaling up whilst improving the effectiveness, we looked into the ImageNet image classification task which has far more data than the public person re-identification datasets. We designed a novel model that automatically builds a hierarchical category structure over the large amount of classes and optimizes the search paths that minimizes the classification time. This model can be directly applied to person re-identification in a large scale.

(2). Making the model adaptive and robust to changing environments

We mainly focused on two related learning problems that can lead to feasible solutions to this objective.

① **Transfer learning**. Adaptation and generalization belongs to the general machine learning problem of transfer learning. We worked on making a deep model that trained on a large dataset effectively transferred to a hard classification problem which only has very small amount of data.

② **Active Learning**. Active learning can help solving the problem when labels are very hard to get, like person re-identification in real scenarios. For pushing the research boundaries, we proposed novel models for two most promising state-of-the-art approaches: the bandit algorithm and the deep reinforcement learning algorithm.

4．研究成果

(1-1) **Dictionary learning**.

As shown in Figure 1, we proposed a novel approach for multiple-shot person re-identification when multiple images or video frames are available for each person, which is usually the case in real applications. Our approach collaboratively learns camera-specific dictionaries and utilizes the efficient l2-norm based collaborative representation for coding, which has shown great superiority in terms of both effectiveness and efficiency to all related existing models.

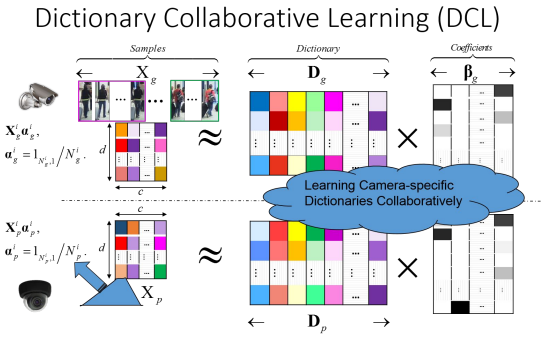

Figure 1: Dictionary Co-learning Model

(1-2) **Metric Learning**.

Towards the challenging variability and sparsity of the data, we proposed NSML (Neighborhood Structure Metric Learning) model (Figure 2) to learn discriminative dissimilarities on a neighborhood structure manifold. Experiments demonstrated the advantage of NSML in terms of effectiveness, robustness, efficiency, stability, and generalizability.

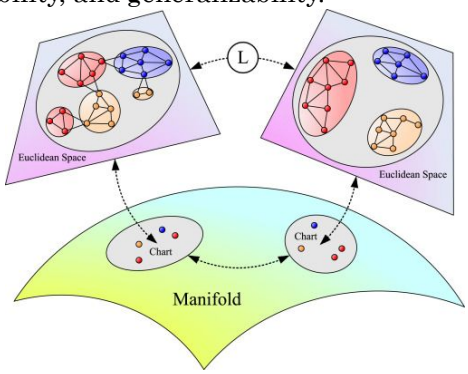

Figure 2. Illustration of the neighborhood structure manifold.

(1-3) **Locality**.
We proposed a novel set-based matching model, "Locality Based Discriminative Measure", to re-identify the human body when a set of test samples for each person are available. A new set-to-set dissimilarity is crafted considering both majorities and minorities of samples from each pair of sets. The discriminability of this dissimilarity is then further exploited by the local metric field; it can thereby serve as a more capable low-level measure to support the high-level measure for the final matching. Extensive experiments on widely used benchmarks demonstrate that our proposal remarkably outperforms state-of-the-arts.

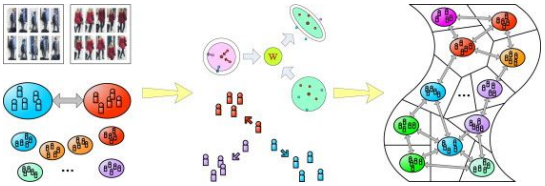

Figure 3. Locality-based discriminative measure (LBDM)

(1-4) **End-to-end deep learning**.
There is a lack of effective way to model the motion information (for example, gait) which can help identifying a person in real scenarios due to the complexity of the data (background, pose changes, occlusions, etc.). As shown in Figure 4, we propose an end-to-end deep learning model with a special component called Temporal Convolution Network (Temporal ConvNet) which does convolution only in the time space (1D convolution). Working together with spatial CNN and normal RNN, it can greatly improve the re-identification performance with almost ignorable additional cost (about only 1% increase in storage and computation).

In greater details, this work has the following novelties.

- ✓ Introduced a new network architecture called T-CN (Temporal-Enhanced Convolutional Network) for video-based person re-identification.
- ✓ Learned low-level and mid-level motion representation which was hard to achieve by existing solutions.
- ✓ Built a whole end-to-end learnable model, which can process videos of various lengths in an efficient way.
- ✓ Showed the superiority of our proposal, in terms of effectiveness, efficiency and generalization ability.

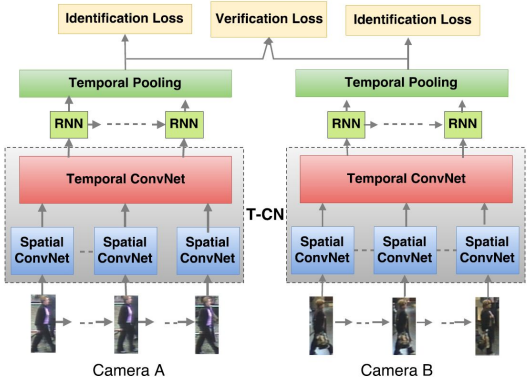

Figure 4. The overall architecture of the proposed Temporal-Enhanced Convolutional Network

(1-5) **Hierarchical structure**.
Due to the lack of publicaly availble datasets for person re-identificaiton, we designed a scalable classification model that can handle a large number of categories and demonstrate its superority on image classification tasks instead. As shown in Figure 5, the proposal is about learning hierarchical interclass structures. Specifically, we first design a fast algorithm to compute the similarity metric between categories, based on which a visual tree is constructed by hierarchical spectral clustering. Using the learned visual tree, a test sample label is efficiently predicted by searching for the best path over the entire tree. The proposed model was extensively evaluated on the ILSVRC2010 and Caltech 256 benchmark datasets for showing its superority.

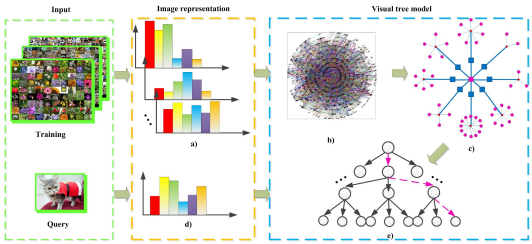

Figure 5. The hierarchical learning framework

(2-1) **Transfer learning**.
Convolutional Neural Networks (CNNs) possess great potential to perform well on many classification tasks, including person re-identification. However, CNNs usually require a large amount of carefully labeled data for training, which is hard to obtain in many real applications. In this paper, we propose a new approach for transferring pre-trained deep networks such as VGG16 on Imagenet to small datasets,

using multi-instance multi-label (MIML) task as an example. We extract features from each group of the network layers and apply multiple binary classifiers to them for multi-label prediction. Moreover, we adopt an L1-norm regularized Logistic Regression (L1LR) to find the most effective features for learning the multi-label classifiers. The experiment results on two most-widely used and relatively small benchmark MIML image datasets demonstrate that the proposed approach can substantially outperform the state-of-the-art algorithms, in terms of all popular performance metrics. The framework our proposed transfer learning model is illustrated in Figure 6.

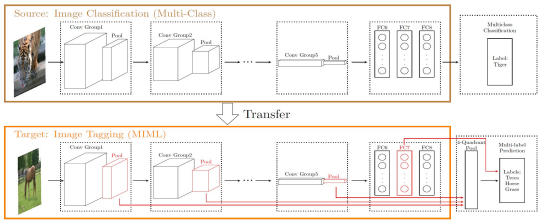

Figure 6. The proposed transfer learning framework

(2-2) **Active Learning**.
Active learning aims to reduce annotation cost by predicting which samples are useful for a human teacher to label. However it has become clear there is no best active learning algorithm. Inspired by various philosophies about what constitutes a good criteria, different algorithms perform well on different datasets. This has motivated research into ensembles of active learners that learn what constitutes a good criteria in a given scenario, typically via multi-armed bandit algorithms. Though algorithm ensembles can lead to better results, they overlook the fact that not only does algorithm efficacy vary across datasets, but also during a single active learning session. That is, the best criteria is non-stationary. This breaks existing algorithms' guarantees and hampers their performance in practice. We propose dynamic ensemble active learning as a more general and promising research direction. We develop a dynamic ensemble active learner based on a non-stationary multi-armed bandit with expert advice algorithm. Our dynamic ensemble selects the right criteria at each step of active learning. It has theoretical guarantees, and shows encouraging results on 13 popular datasets.

Besides exploring the bandit algorithm, we also propose to treat active learning algorithm design as a meta-learning problem and learn the best criterion from data. We model an active learning algorithm as a deep neural network that inputs the base learner state and the unlabelled point set and predicts the best point to annotate next. Training this active query policy network with reinforcement learning, produces the best non-myopic policy for a given dataset. The key challenge in achieving a general solution to AL then becomes that of learner generalisation, particularly across heterogeneous datasets. We propose a multi-task dataset-embedding approach that allows dataset-agnostic active learners to be trained. Our evaluation shows that AL algorithms trained in this way can directly generalise across diverse problems.

All the three machine learning models provide new possiblities for make the person re-identification model work under the transfer/active learning condition so that the great efforts of manual annotation can be minimized.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　3　件）

[1] Yanyun Qu, Li Lin, Fumin Shen, Chang Lu, Yang Wu, Yuan Xie, Dacheng Tao, "Joint Hierarchical Category Structure Learning and Large Scale Image Classification", IEEE Transactions on Image Processing, vol.26, no.9, pp. 4331-4346, 2017.**査読有**.
DOI: 10.1109/TIP.2016.2615423

[2] Wei Li, Yang Wu, Jianqing Li, "Re-identification by Neighborhood Structure Metric Learning", Pattern Recognition. 61, pp. 327-338, 2017. **査読有**.
DOI: 10.1016/j.patcog.2016.08.001.

[3] Wei Li, Yang Wu, Masayuki Mukunoki, Yinghui Kuang, Michihiko Minoh, "Locality Based Discriminative Measure for Multiple-shot Human Re-identification", Neurocomputing, Volume 167, 1 November 2015, pp. 280-289, 2015. **査読有**.
DOI: 10.1016/j.neucom.2015.04.068

〔学会発表〕（計　7　件）

[1] Yang Wu, Jie Qiu, Jun Takamatsu,

Tsukasa Ogasawara. “Temporal-Enhanced Convolutional Network for Person Re-identification”. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), New Orleans, USA, 2018. pp. 7412-7419. **査読有**.

[2] Kunkun Pang, Mingzhi Dong, Yang Wu, Timothy Hospedales, “Dynamic Ensemble Active Learning: A Non-Stationary Bandit with Expert Advice”, 24th International Conf. on Pattern Recognition (ICPR), 2018. **査読有**.

[3] Kunkun Pang, Mingzhi Dong, Yang Wu, Timothy Hospedales, “Meta-Learning Transferable Active Learning Policies by Deep Reinforcement Learning”, International Workshop on Automatic Machine Learning, Collocated with the Federated AI Meeting (ICML, IJCAI, AMAS, and ICCBR), Stockholm, Sweden, 2018. **査読有**.

[4] Yang Wu, Jie Qiu, Jun Takamatsu, Tsukasa Ogasawara. “Temporal-Enhanced Convolutional Network for Person Re-identification”. The 12th International Workshop on Robust Computer Vision, Nara, Japan, 2018. **査読無**.

[5] Mingzhi Dong, Kunkun Pang, Yang Wu, Jing-Hao Xue, Timothy Hospedales, Tsukasa Ogasawara. “Transferring CNNs to Multi-instance Multi-label Classification on Small Datasets”. IEEE International Conference on Image Processing, Beijing, China, 2017. pp. 1332-1336. **査読有**. DOI: 10.1109/ICIP.2017.8296498

[6] Yang Wu, Dong Yang, Ru Zhou, Dong Wang. “Dictionary Co-learning for Multiple-shot Person Re-identification”. The 11th Chinese Conference on Biometric Recognition, Chengdu, China, 2016, pp. 675-685.**査読有**.
DOI: 10.1007/978-3-319-46654-5_74

[7] Chang Lu, Yanyun Qu, Jianping Fan, Yang Wu, Hanzi Wang. “Hierarchical Learning for Large-scale Image Classification via CNN and Maximum Confidence Path”. In Proc. of the 16th Pacific-Rim Conference on Multimedia, Gwangju, Korea, 2015, pp. 236-245. **査読有**.
DOI: 10.1007/978-3-319-24078-7_23

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
伍 洋（WU, Yang）
奈良先端科学技術大学院大学・研究推進機構・特任助教
研究者番号：30750559