6版

様　式　C－19、F－19－1、Z－19　（共通）

科研費 KAKENHI

科学研究費助成事業　　研究成果報告書

令和　3 年　5 月 27 日現在

機関番号：１４４０１

研究種目：研究活動スタート支援

研究期間：2019～2020

課題番号：１９Ｋ２４３６４

研究課題名（和文）Toward a Multi-Gait Analysis/Recognition System

研究課題名（英文）Toward a Multi-Gait Analysis/Recognition System

研究代表者

ＡＬＬＡＭ　ＳＨＥＨＡＴＡ・ＡＬＬＡＭ（ALLAM, ALLAM）

大阪大学・産業科学研究所・特任研究員（常勤）

研究者番号：７０８５０７６７

交付決定額（研究期間全体）：（直接経費）　　400,000円

研究成果の概要（和文）：この研究プロジェクトを通じて、私は次のタスクを完了しました。
1-182人の歩行被験者を含む単一の歩行データセットを作成します。 2-グループ歩行実験のラボ記録からのマルチ歩行データセットのコンパイル。 3-シングル歩行データセットとマルチ歩行データセットの両方のオプティカルフローと高密度軌道を抽出します。 4-オプティカルフローと密な軌道を使用して、抽出された各密な軌道とともにオプティカルフローの動き情報の集約を検討することにより、歩行対象ごとにローカルモーション記述子を作成しました。 また、各軌道の相対位置と形状記述子を計算しました。 5-計算されたローカル記述子を指定して、各サブジェクト.

研究成果の学術的意義や社会的意義
I introduced the first steps toward a multi-gait recognition system. Ihave prepared the first multi-gait dataset for research purposes. The dataset contains, videos, optical flow, BBXs,and dense trajectories for all subjects.I introduced a robust motion feature representation for model evaluation.

研究成果の概要（英文）：Through our this research project, I finished the following tasks:
1- Creating the single gait dataset which containing 182 walking subjects. 2- Compiling the multi-gait dataset from my lab recordings of group walking experiment. 3- Extract the optical flow and dense trajectories of both the single and multi gait dataset. 4- I used the optical flow and dense trajectory to build local motion descriptors for each walking subject by considering the aggregation of the optical flow motion information along with each extracted dense trajectory. As well, I computed the relative position and shape descriptor for each trajectory. 5- To build a global motion descriptor for each subject given the calculated local descriptor, I applied the robust fisher vector encoding technique. To evaluate the proposed feature representation, I measured the pairwise similarity between single and multi global descriptor for each subject. I used it as well in gait relative experiment.

研究分野：Gait modeling

キーワード：Gait recognition　Multi objects tracking　Optical flow　Dense Trajectory　FIsher vecotr encoding

1.　研究開始当初の背景

The human gait modality is adopted to express the walking style of a moving person. This modality is mainly used in video surveillance to capture the person's identity. The traditional gait recognition approaches have been applied to the single-person walking data. However, in real environments, a person often walks with other people. This introduces challenges for such recognition systems. By assuming the multi persons walking scenario, capturing multi-gait GEI for each individual becomes a challenging task. Developing the gait-based analysis/recognition systems has become an attractive topic in computer vision field. Such systems have been used in security surveillance applications, such as the authentication and suspicious behaviors detection [Xin Chen et al. 2018]. The baseline stage of existing gait analysis/recognition systems is the computation of binary silhouettes from the video of the walking person. This binary silhouette-based descriptor is called gait energy image (GEI) and it is usually captured by temporal averaging of the binary silhouette sequence [Han and Bhanu, 2006].
However, such descriptor suffers from the limitations of dynamic backgrounds and non-static camera situations. On the other hand, most of current gait analysis/recognition approaches adopt the single person walking scenario. Those approaches have achieved satisfactory performances with a good tolerance for different covariates (i.e. cloths and carrying objects). However, in real environments, a person often walks with other people. This introduces challenges for such recognition systems. I claim that it is more practical to tackle multi-gait recognition in crowded low-density scenes. Based on the success of local feature descriptors in action and gait recognition systems, particularly those descriptors developed from dense trajectories, I expect it to serve well in the multi gait problem instead of GEI.

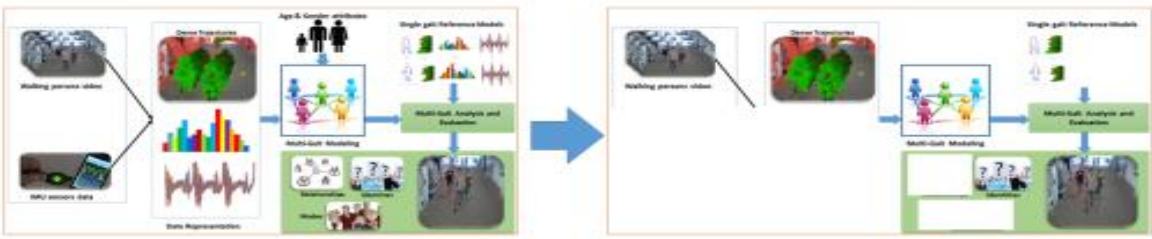

Fig. 1: The block diagram of the proposed model: slightly adapted after the budget reduction (left)

(a) Relative Position of DT　　(b) Multi-gait/Single-gait dataset

Fig. 2: (a) The DT relative position, (b) Sample frames from the multi-gait/ sequences (right) versus the single-gait sequences (left).

2. 研究の目的

The ultimate goal of this project was to initiate a baseline smart camera-based surveillance system for Multi-Gait recognition in sparse to medium-crowded scenes. The target system will automatically analyze the video data captured by fixed surveillance cameras and autonomously identify the person's identity and behavior based on his/her gait while walking with/through multiple persons. I proposed the framework structure to be introduced into the following phases:
Offline Phase: Given a video recorded for a particular person walking alone for a period of time,

- The dense trajectories for this walking person are obtained then refined by removing the background and shadow trajectories.
- A novel descriptor will be computed for the gait features involved in the refined trajectories. These gait features are used to train a robust classifier.
- Extract the optical flow information for each subject's video.
- Extract the binary silhouette sequences for each subject for trajectories refinement prepossessing step.

Online Phase: Given an online video record, the person of interest walking alone or with other individuals, do the following:

- Obtain the dense trajectories from the video stream, then refine them by removing the
- background and shadow trajectories.
- A novel trajectories association method will be proposed based on the individuality
- feature that I will be extracted from the neighbor trajectories.
- Extract the optical flow information of the group walking video.
- Extract foreground pixels and Bounding boxes (BBxs) of the group walking video for
- trajectories refinement prepossessing step.
- Using each subject's BBxs sequences, and the foreground pixels, I extract the corresponding individuals' detection, their optical flow information through the frame sequence, as well as the refined associated dense trajectories.

３．研究の方法

Actually, collecting the data for this project was a challenge. In the project budget distribution, I assigned a good amount of money for data collection. Although JSPS reduced the project's budget from 3 million yen to 500,000 yen, I decided to find an alternative. After consultation with my lab director, I decided to compile my dataset from the available data in my lab. Particularly, I compiled the multi-gait dataset from the recorded videos in my lab. It was challenging to find the proper recordings that match with my problem setting. As shown in Fig. 2, I tried to find the video recordings that contain the same subjects appear in both the multi and single gait scenarios. To achieve this, I collected 182 videos for the walking subjects while they are walking individually, as shown in Fig. 2(b) right (Single-gait dataset). For the multi-gait dataset, I collected the recording video that contains as many subjects as possible walking together. In total, around 80 subjects from the single-gait dataset have appeared in the multi-gait video.

４．研究成果

1) Dense trajectories extraction

Using the method proposed in [6] I extracted the dense trajectories (DTs) of the walking videos. Dense trajectories are extracted for multiple spatial scales where feature points are sampled on a grid spaced by a grid of pixels and tracked in each scale separately. Each extracted trajectory t j is represented by a sequence of n (x_y) coordinates as follow

$$t_j = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\} \tag{1}$$

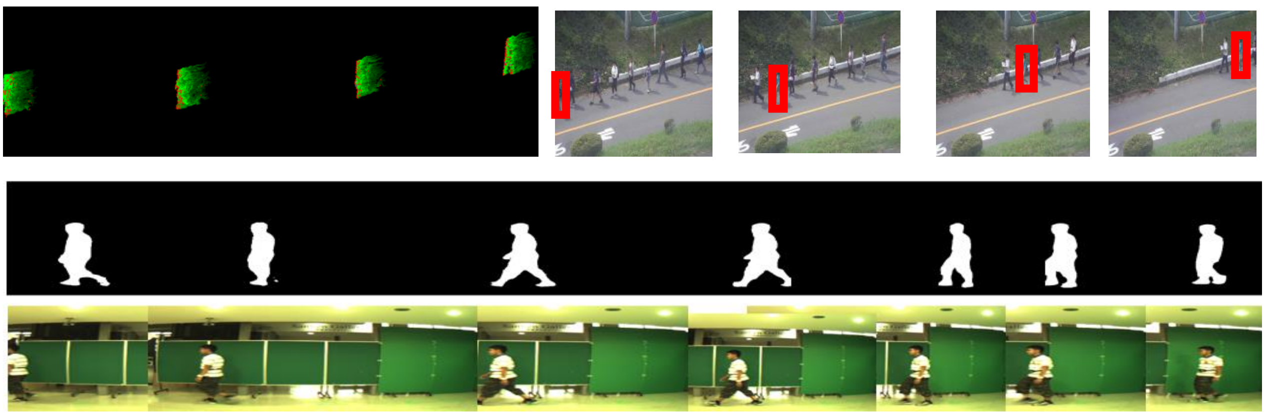

Fig.3. Dense trajectory/ BBXs extraction (top), Foreground masks extraction

2) Building the motion feature descriptor

This trajectory encodes the local motion information of the body parts of the walking person. I had a claim that each body part has a contribution to express the person's identity. Therefore, I proposed to relate each extracted DT to its corresponding body parts by computing

1. The relative position considering the centroid of the boundingbox of each detection.
2. The shape descriptor.
3. The average motion magnitude.
4. The motion orientation.
5. HOF (i.e., histogram of optical flow) descriptor for each trajectory that encodes the local dynamics along with each trajectory.

Relative position: Suppose that the centroid point of the bounding box of the detected subject is denoted as

p^0 has a coordinate (x_0, y_0). Let a point i lies on trajectory j is denoted as pji has a coordinate (x_i, y_i). (see Fig. 2). I can define the relative position of this point as

$$\hat{p}_i^j = p_i^j - p_0, \left(\hat{x}_i^j, \hat{y}_i^j\right) = \left(x_i^j, y_i^j\right) - (x_i, y_i) \quad (2)$$

I hence can define the relative position of a trajectory $t^j = \left\{\hat{p}_i^j\right\}^{i=1:L} \in \mathbb{R}^{15}$ and L is the length of the trajectory (in my experiments I set L = 15 frames).

The shape descriptor: The shape of a trajectory encodes local motion patterns. Given a trajectory of length L, I can describe its shape by a sequence displacement vector And I adopted $\Delta p \in \mathbb{R}^{28}$

$$\triangle P = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t) \quad (3)$$

The average velocity and motion orientation: To enrich the representation of the local descriptor of each trajectory, I considered the computation of motion orientation and velocity of each trajectory. Given the trajectory t^j has x − y points locations $t^j = \{x_i, y_i\}\ i = 1{:}L$. I computed the velocity and motion orientation respectively as

$\bar{V} = \frac{1}{L}\sum_{i=1}^{L}\left\|\vec{V_i})\right\|$, $\bar{\theta} = \frac{1}{L}\sum_{i=1}^{L} tan^{-1}\left(\frac{v_i}{u_i}\right)$ where $\left\|\vec{V_i})\right\| = \sqrt{u_i^2 + v_i^2}$ and $u_i = \frac{\partial x_i}{\partial t}$, and $v_i = \frac{\partial y_i}{\partial t}$ and both $\bar{V}$ and $\bar{\theta}$ are $\in \mathbb{R}$.

HOF descriptor of dense trajectory: HOF descriptors have become widely used in many recognition and video classification approaches [4,5,6,2]. I adopted it to encode the local motion information for each DTs. Following [6], I computed the HOF descriptor from the 3D spatial-temporal volume around the trajectory based on the default setting of HOF computation. As a result, the global HOF descriptor f for each trajectory is concatenated over the sub-block with dimension of 12 × 9bins = 108 and f ∈ R108. It worth noting that I used the recent deep SpyNet-Flow extractor [3] to get the optical flow information of the walking video. To remove the background and data clutter optical flow, I used the extracted foreground masks.
for this task and also for trajectories refinements. I used the recent deep model, nicked-named as PointRend [1], for image segmentation to extract the foreground masks as it produces high- quality foreground masks for Multi-gait and Single-gait datasets as shown in Fig. 2 and Fig. 3. In total The final computed descriptor h j of trajectory j is introduced as the concatenation of the above computed descriptors as

$$h_j = \left\{F^j, t^j, \hat{p}_i^j, \bar{\theta}, \bar{V}\right\}$$

Fisher Vector Encoding: The FV mechanism is a well-known encoding method used to represent a set of local descriptors into a single rich global descriptor. FV encoding of the local D-dimensional descriptors is
based on a trained generative model, such as the Gaussian mixture model (GMM). As I computed several HOFs (typically 3,000–5,000) for each walking video, I used FV to encode these HOFs into one global FV. The ultimate goal is leveraging the FVs mechanism to encode the gait motion information from the walking video into a single global descriptor used to train the classifier.

Disclaimer and future work: According to the budget distribution in my laboratory, I had to pay only 10% (around 8 hours per month) of my efforts to this project and 90% to another project. As a result, I have finished building the motion descriptors for the single-gait dataset. I am still working on the multi-gait dataset motion descriptors. The dataset is confidential, and there is no permission yet to release it to be publicly available. BTW, I have examined my proposed motion feature representation in gait relative attribute estimation work, and it achieved a good performance in terms of classification accuracy compared with the deep- learning features (VGG16), as shown in Table 1. I expect better performance for the multi-gait recognition task. One of my future directions is to benefits from the available high-quality foreground masks of the proposed dataset. I propose to treat the single gait-dataset foreground masks as binary silhouette sequences and use them for training a deep learning model Once trained, I will use the multi-gait silhouette to examine the model performance.

GRA	General goodness	Stately	Cool	Relax	Arm swing	Step length	Walking speed	Spine	Accuracy (avg.)
GEI+VGG+R-SVM	67.70	66.7	58.00	71.3	66.3	61.9	57.3	63.1	64.04
DTs+FV+R-SVM	60.2	42.86	***62.89***	***75***	***68.75***	**75**	**76.77**	66.67	66.02

Table 1: Comparison of the proposed Dense trajectory-based representation with benchmark in classification accuracy

References

1) Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020)
2) Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies (2008)
3) Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4161–4170 (2017)
4) Uijlings, J., Duta, I.C., Sangineto, E., Sebe, N.: Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. International Journal of Multimedia Information

Retrieval 4(1), 33–44 (2015)
5) Uijlings, J.R., Duta, I.C., Rostamzadeh, N., Sebe, N.: Realtime video classification using dense hof/hog. In: Proceedings of international conference on multimedia retrieval. p. 145. ACM (2014)
6) Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 3169–3176. IEEE (2011).

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

During this project, Prof. Gomaa has attended several online meeting as a co-advisor. He contributed by his valuable comments that helped me in my research.

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	槇原　靖 (Makihara Yasushi) (90403005)	大阪大学・高等共創研究院・教授 (14401)	
研究協力者	八木　康史 (YAGI Yasushi) (60231643)	大阪大学・産業科学研究所・教授 (14401)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関