View-invariant Feature using Pose Information and Flexible Matching Algorithm for Action Retrieval

Noboru Yoshida and Jianquan Liu
Biometrics Research Laboratories, NEC Corporation, Japan
E-mail: {n-yoshida14, jqliu}@nec.com

***Abstract*—Action retrieval and detection utilizing pose information has two difficulties in industrial applications, the occlusion and variation of human actions. To overcome these difficulties, we propose a new normalization method to generate view-invariant pose feature that is corresponding to the key-points of the original pose information, and design flexible matching algorithms to perform high-accurate action retrieval. The proposed methods enable users to easily and flexibly perform action retrieval with weight constraints on specific body parts or neglect of invisible body parts for similarity computation. The experimental results are reported to show that our feature and flexible matching algorithm outperforms the state-of-the-art methods in a simulated dataset with annotated multi-view 2D poses and a real-world video dataset.**

I. Introduction

Human action detection (or recognition) has been studied for over two decades [1]. As far as we know, the demand of applications has a diverse range, such as detecting unsafe actions (e.g. fall down, crouch down) and persons who need support (e.g. wheel chair users) in public places, the editing or the analysis on a large amount of video archives based on human actions. A number of existing methods adopted traditional machine-learning techniques [2], [3], [4], [5], [6], and especially in recent years, deep-learning techniques [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] to handle specific action detection.

However, to perform robust and generic detection, the deep-learning methods have to be fine-tuned and well trained to extract invariant deep features against the change of following factors: 1) outward appearance (cloth, accessory, belongings, etc.), 2) shooting environment (light or sunlight condition, background, etc.), and 3) camera parameters (position, direction, focal length, etc.). Therefore, the training process inevitably becomes highly time- and cost-consuming to cover these variations requiring a large amount of training data with annotations.

To solve this challenging problem, researchers are trying to abandon these learning-based methods, alternatively adopt retrieval approaches [17], [18], [19], [20], [21], [22], [23], [24], [25] using features of optical flow, pattern histogram, etc. to handle action detection in the manner of query matching. Especially human pose information extracted from image is promising feature for human action detection and retrieval [22], [23] because, this feature overcome the changes of (1) outward appearance and (2) shooting environment. Furthermore, learning view-invariant embedding space by training the pair of 3D pose and the projected multi-view 2D pose to overcome the changes of (3) camera parameters. This technology was applied to not only view-invariant similar pose retrieval in images, but also view-invariant similar action retrieval in videos [24].

However we have to overcome another two difficulties to achieve high performance in industrial applications. The first is the variation of poses and their movements regarding the same action. For example, the pose of sitting on the ground with/without crossing arms. Both are "sitting on the ground" but different poses from each other. In this case, conventional method cannot retrieve one by utilizing the other as the given query, because the method treat whole body information equivalently and retrieve the "same pose" as the given query. To tackle this problem, we apply weight constraints on specific body parts (i.e., key-points) for similarity computation to retrieve similar actions.

The second difficulty is occlusion. A part of estimated pose information is often missed due to the overlapping between human and human, or human and object in real-world scenarios. In the conventional techniques, the missing pose key-points are complemented for full-pose matching based on the time series information and the information of neighbor key-points. However the reliability of the complemented information is limited and insufficient. To achieve high performance in this situation, partial-pose matching only on visible body parts are considered as more feasible.

Similarly inspired by the related work, we utilize pose information and transform them to view-invariant feature, and adopt retrieval approach to perform high-accurate human action retrieval. To overcome the above-mentioned two challenges, we propose a new normalization method to generate view-invariant pose feature that is corresponding to the key-point of the original pose information, and design flexible matching algorithms for action retrieval. The proposed methods enable users to easily and flexibly perform action retrieval with weight constraints on specific body parts or neglect of invisible body parts for similarity computation.

We implemented the normalized pose feature and the matching algorithm in action retrieval system that can retrieve similar poses or actions by giving an image or a few second sample video with specified action (Fig. 1). Evaluation was done by using annotated multi-viewpoint 2D pose dataset simulated by motion capture data, and UT-kinect [26] that is widely used human action video dataset. The experimental results

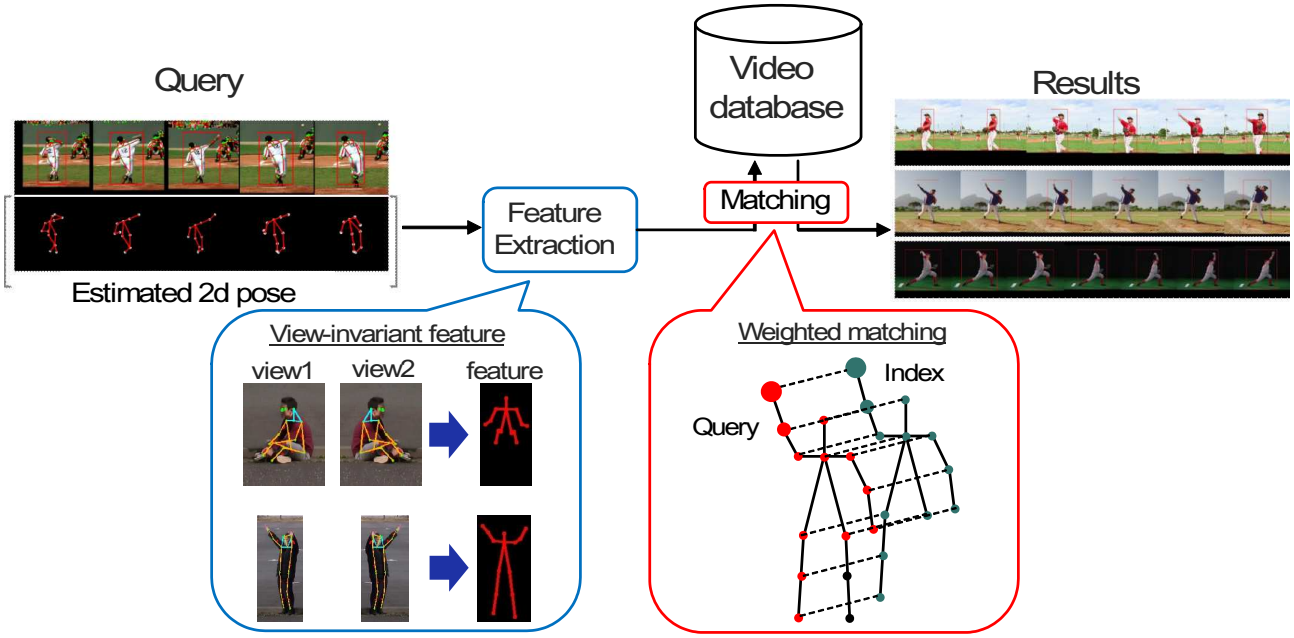

Fig. 1. The overview of our method. Key-ideas are view-invariant pose feature, and weighted matching algorithm in which missing key-points are ignored and characteristic key-points associated with weights are indicating by bigger points.

are reported to show that our approach demonstrates higher precision and recall than the state-of-the-art methods.

II. Related works

A. 2D pose estimation

In recent years, many pose estimation technologies have been proposed for action recognition [27], [28], [29], human tracking [30], [31], and human re-identifications [32]. Pose information is human skeleton structure with 18 key-points, and invariant feature against outward appearance and shoot environment.

There are two major approaches, top-down [33], [34] and bottom-up [35], [36], [37], [38]. The top-down approach embeds a human detector at the beginning of its data processing unit, and detects each key-point in the detected human bounding box. However, as in real-world surveillance where the environment is often crowded and a part of human body is often occluded, thus human detector tends to fail. Such a problem has been pointed out by Gkioxari et al. in their research [39].

To solve this real problem in industrial applications, compared to the top-down method, the alternative approach called bottom-up method is more robust for human and action recognition, which constructs human pose structure based on recognized key-points. Such a bottom-up method first recognizes human key-points on visible area of human body in the whole image, then associates those visible key-points into individual persons and generates human bounding boxes. As a result, human bounding boxes may not enclose a person's full body but the detection itself is more reasonable and flexible.

In our work, we applied bottom-up pose estimation method to achieve high-accurate action retrieval robust to occlusion.

B. Pose retrieval

The technology of similar image retrieval is proposed to use 2D human pose information for similarity matching between the data and the given query specified by a user [22], [23], [25]. To trigger the process of pose retrieval, there are different ways to specify a query with pose information. For instance, manually operate the pre-installed pose structure by a user interface [23], input the Kinect sensor data [23], use the output of pose estimation engine mentioned above [22], [23], or input a user-written sketch [25] to trigger a retrieval task.

However, the retrieval performance drastically decreases because the recognized 2D pose information often consequently changes when camera parameters (position and angle) or person's orientation change in different environments. To overcome this problem, learning view-invariant embedding space by training the pair of 3D pose and the projected multi view 2D pose. The matching phase uses a general method of measuring the L_p norm between embedded features [24]. However, in industrial applications, the retrieval performance might degrade by the following two difficulties.

The first is the variation of poses regarding the same action. Conventional features and matching algorithm treat the whole body information equivalently so that the same action yet different pose is hard to be retrieved. To overcome this difficulty, we propose an approach of retrieval with weight constraints on specific body parts or neglect of invisible body parts for similarity computation.

The second is occlusion. A part of estimated pose information is often missed due to the overlapping of human-human and human-object in real world scenarios. Conventional methods complement the missing key-points based on the time series information and the information of neighbor key-points in the preprocessing phase. These methods need 2D full-pose as input of deep networks. However the reliability of the complemented information is limited and insufficient. To achieve high performance in the occlusion situation, partial-pose matching only on visible body parts are considered as more feasible.

To solve the above-mentioned two challenging difficulties in industrial applications, we propose a new normalization method to generate view-invariant pose feature that is corresponding to the key-point of the original pose information, and design flexible matching algorithms for action retrieval. The proposed methods enable users to easily and flexibly perform action retrieval with weight constraints on specific body parts or neglect of invisible body parts for similarity computation.

C. Action retrieval

An action retrieval methods by using a time-series pose feature as a query have been proposed [18], [25]. Most of researches utilize 3D human pose information obtained by Kinect sensors, magnetic sensors, and motion capture system. On the other hand, Sun et al. utilize only 2D pose information as input and transfer it to view-invariant embedding feature for view-invariant action retrieval task in videos [24].

In these related works, it is a common way to match frames between two videos by using Dynamic Time Warping (DTW) [40] method and to measure the distance between videos by the sum of distances between corresponding frames.

In our work, we also apply the proposed feature and matching algorithm to action retrieval task in videos. Frame matching was performed by DTW as in the existing research, and the proposed matching algorithm considering weights and defects was used to calculate the frame-frame distance.

III. Proposed Method

A. View-invariant feature

As mentioned before, we also utilize pose estimation technique to extract pose skeleton structure with 18 key-points and convert it into view-invariant feature. Existing method [24] also proposed the view-invariant embedding feature obtained by deep metric learning. With this method, the obtained features lose the original key-point label (such as shoulders, elbows, etc.) information.

Assume the situation that retrieving the people who sit on the ground in real world surveillance, sitting pose would varies from person to person. For example, one is crossing arms, and the other is not crossing arms as shown in Fig. 2 (but both are sitting on the ground). In this case, conventional method cannot retrieve one by utilizing the other as query, because the method treat whole body information equivalently and retrieve the "same pose" as query. However in this case, the system users would easily come up with the idea to more weight on the leg shape similarity or ignore the upper body information. To realize this idea easily, the view-invariant features should hold the original key-point label. Based on this policy, our feature is designed as each dimension has a one-to-one correspondence relationship with each key-point of the original pose information.

This approach has another advantage in treating missing key-points. Conventional methods complement the missing key-point based on the time series information and the surrounding key-point information in preprocessing phase because the whole pose information is needed as an input of deep networks. However, the reliability of the complemented information is limited and insufficient. On the other hand, our feature extraction method does not need the missing key-points complementation and the feature holds the "missing" information in each dimension. So the higher performance is achieved by matching only with the visible body part information.

We assume that almost all cameras are installed with the vertical axis in the screen parallel to the direction of gravity in the real world. Based on such an assumption, the y-axis of the pose structure estimated from the image is almost invariant with respect to the pan angle of the camera. Then, we can normalize the y coordinate of a pose proportional to the person's height (pixel) in the screen, which changes with the tilt angle and focal length. By this idea, we can generate a normalized pose feature that is invariant to the change of camera parameters by the following equation.

$$f_i = \frac{pose\_y_i - core\_y}{p\_height},$$

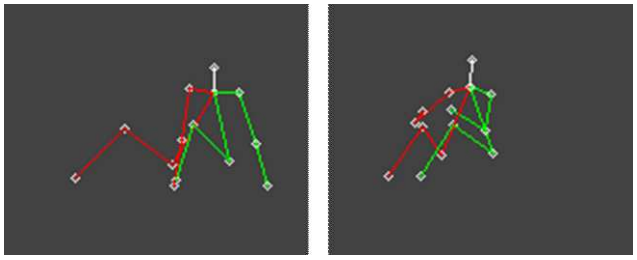

Fig. 2. Poses of sitting on the ground without crossing arms (left) and with crossing arms (right).

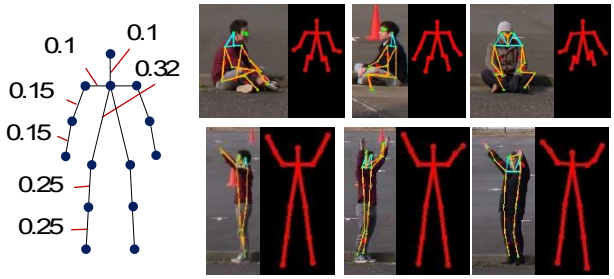

Fig. 3. Standard human model (left) and pair images of estimated 2d pose and visualized feature extracted by proposed method (right).

where $pose\_y_i$ is y component of ith 2D key-point, $core\_y$ denotes the neck key-point, and p_height is the estimated height of a person in the screen. If ith key-point is missed, f_i set to -1. p_height is also estimated by 2D pose information.

Since 2D pose information frequently has missing key-points due to occlusion, we propose a robust height estimation method for key-points missing. To achieve this, we define a standard human model (Fig. 3) which represent the relationship between the length of each part (such as limbs and shoulder length) and height, and estimate the height from each part by applying to the model. From these calculated heights, those with larger values are extracted and averaged to be the estimated height.

Note that the each dimension of the feature has a one-to-one correspondence with each original pose key-point, and the feature hold "missing key-point" information in each dimension.

The pair of 2D pose and visualized proposed feature is also shown in Fig. 3. In visualization, the x component was set to a constant value, and the y component is changed according to the proposed feature value. It can be confirmed that the proposed feature is invariant with respect to the clothes, background and orientation of the person.

B. Frame Matching function

For the distance function between features, the following formula which applies the general $L1$ distance was used.

$$L(f_a, f_b) = \frac{1}{n}\sum_{i=1}^{n} |f_{a,i} - f_{b,i}|,$$

where f_i is ith dimension of feature vector, and n is the number of key-points.

When a part of the key-points is missing, matching by using only the visible key-point information is useful. Therefore, when the key-point is missing, the following distance function is used.

$$L\_lack(f_a, f_b) = \frac{\sum_{i=1}^{n} |f_{a,i} - f_{b,i}| \times l_i}{\sum_{i=1}^{n} l_i},$$

$$l_i = \begin{cases} 1 \ (f_{a,i} \neq -1 \ and \ f_{b,i} \neq -1) \\ 0 \ (f_{a,i} = -1 \ or \ f_{b,i} = -1). \end{cases}$$

Furthermore, in order to retrieve various actions that differ from person to person with high accuracy, weight the similarity of characteristic body parts is effective. The weighted distance function by this method is given by

$$L\_lack\_weight(f_a, f_b) = \frac{\sum_{i=1}^{n} |f_{a,i} - f_{b,i}| \times W_i}{\sum_{i=1}^{n} W_i},$$

$$W_i = \begin{cases} w_i \ (f_{a,i} \neq -1 \ and \ f_{b,i} \neq -1) \\ 0 \ \ (f_{a,i} = -1 \ or \ f_{b,i} = -1), \end{cases}$$

where w_i is weight on the ith feature. However, it is difficult for the user to manually weight each query. To solve this problem, we define the upright state (Fig. 3 left) as the reference pose and propose an algorithm that weights the each key-point based on the difference from the reference pose. w_i is calculated by

$$w_i = \frac{|f_{q,i} - f_{r,i}|}{sum\_length_{r,i}},$$

where f_r is the feature of reference pose, and $sum\_length_{r,i}$ denotes the sum of length between ith key-point of reference pose and neck key-point of reference pose (for example, $sum\_length_{r,Rhand}$ is 0.4, calculated by sum of the length of Neck-Rshoulder (0.1), Rshoulder-Relbow (0.15), Relbow-Rhand (0.15)). In this method, each weight is roughly normalized to 1 to 3.

C. Scene matching algorithm

Matching distance between two features was defined as in the section B. Given the matching distance, we use standard DTW algorithm [40] to align two action sequences by minimizing the sum of frame matching distances, and sum of frame distances was used as the distance between two actions.

IV. Experiment and evaluation

We demonstrate the performance of our feature and matching algorithm through pose retrieval across different camera views. We further show our method can be directly applied to downstream tasks, such as action recognition, without any training.

A. Dataset

For pose retrieval experiments, we validate on an annotated multi-viewpoint 2D pose dataset simulated by motion capture data. Additionally, we also evaluate our method for action retrieval task on UT-Kinect dataset [26].

Motion Capture dataset

We captured several actions using motion capture system in order to obtain the time-series annotated 3D human pose dataset. Then these data are converted into 2D pose by simulating various camera parameters setting to construct a multi-view and annotated 2D human pose dataset. The dataset contains 8 actions that are sit on chair, sit on the ground, lie on the ground, raise right hand, raise left hand, raise both hands, raise right leg, and raise left leg, acted by 10 people.

The obtained 3D pose information is projected to 2D pose by setting camera on a hemisphere (radius = 5m) which centered on the feet of the 3D pose as shown in Fig. 4. Pan and tilt angle was set from 0° to 350° and from 0° to 40° in 10° increments respectively. Thus, a total of 180 types of 2D pose generated from one 3D pose.

UT-kinect

This dataset contains 200 trimmed videos for 10 actions. There are 10 actors and they act each actions two times. Camera parameter is fixed but orientation of actors are changed. We estimate 2D pose key-point from whole image using NeoPose [35] and associated with specific person by the provided person bounding boxes in each frame.

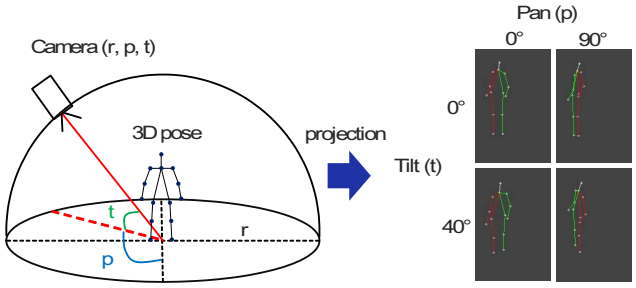

Fig. 4. 2D projection from 3D pose.

B. Action retrieval by pose

Given motion capture dataset, we query using projected 2D key-points from one camera view (pan = 0°, tilt = 0°) and find the nearest neighbors using normal matching distance L (see Section III-B) in the feature space from a dataset (except query pose). In retrieval phase, the dataset is re-organized into three subsets to examine the efficacy of our approach in different conditions of a given query, of which the settings include

(a) the data containing the same person as the given query.

(b) the data using the same camera parameters as the given query (i.e., pan = 0° , tilt = 0°).

(c) all data including the variations of all persons and all camera parameters.

We iterate each pose of all actions and all persons in the dataset as a given query. The experimental results are reported in the average of all queries.

C. Evaluation Procedure

We report the Recall@Precision = 90% for the tasks of pose retrieval, which is the percentage of correct data retrieved from the ground-trues in whole dataset. A retrieval result is considered correct if the action label of the retrieved pose is the same as the query.

D. Baseline Approaches

We compare our method with full-pose matching method which utilizes height normalized 2D pose as the feature [23], and view-invariant deep feature [24] that is trained by Human3.6M dataset without additional pre-training on our own dataset. For fair comparison, the same matching distance L (see Section III-B) is used for the three methods (pose, pr-vipe, and ours).

E. Without occlusion

The result of recall@precision = 90% is reported in Table I. Comparing the top-10 images of pose [23] and ours regarding the setting of sub-set a, and pr-vipe [24] and ours regarding the setting of sub-set a are shown in Fig. 5 and Fig. 6 respectively.

Regarding the setting a, the top-10 images of ours include various orientations of images, indicating our method has better performance than pose due to the view-invariance of the feature. However the performance number is less than pr-vipe. On the other hand, our method outperforms both baselines in setting b, indicating our method is better for robust retrieval on diverse pose. The result of different people can be seen in top-10 images of ours.

In addition, our method returns the best results in the setting c, which comprehensively evaluates the view-invariance and the robustness to the pose variations.

TABLE I
COMPARISON OF ACTION RETRIEVAL RESULT ON MOTION CAPTURE DATA WITHOUT ANY OCCLUSION.

	Recall (%) @ precision = 90%		
Experimental setting	A	B	C
pose [23]	31.5	84.0	19.6
pr-vipe [24]	**97.6**	49.7	49.0
ours	87.7	**91.0**	**72.4**

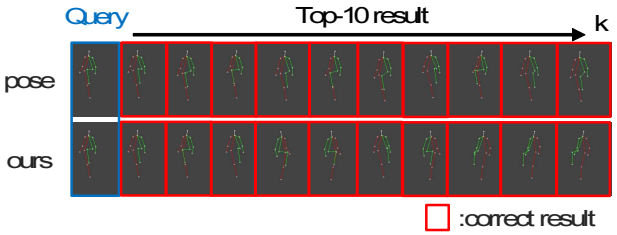

Fig. 5. Comparing the top-10 images of pose and ours in setting a. The query action is raise left leg.

F. With occlusion

In order to evaluate the robustness to the key-point missing, we assume the specific key-point is missed in both query and index poses. We randomly dropped out the one key-point at the end of body (right hand, left hand, head, right foot and left foot) that are frequently missed in real situation. And we controlled the query and index has different key-point missing. Missing key-point is placed to the center of the human bounding box.

In addition to the normal matching distance L, we used L_lack for our features (ours_lack) to retrieve ignoring the missing key-points (see Section III-B). Table II shows the normal matching function is not suitable for retrieval with missing key-point regardless of the type of feature. One way to improve the performance is changing the key-point complement method, however the reliability of the information is limited and insufficient. On the other hand, ours_lack outperform all the other method and only a slight decrease in recall compared to the evaluation without any key-point missing. This result indicate the matching only with visible key-point information is better way for retrieval poses with occlusion.

G. Action retrieval by movie

As a dataset, UT-Kinect [26] was used. As a baseline, we also selected two method, pose [23] and pr-vipe [24]. Matching distance was defined as the L (see Section III-B) between two features. L_lack and L_lack_weight are also

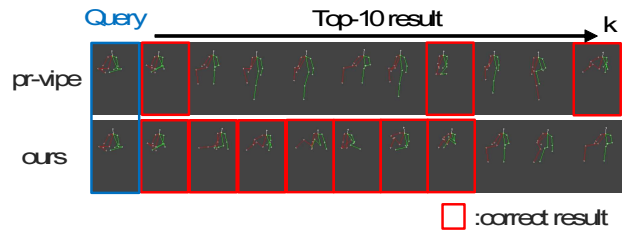

Fig. 6. Comparing the top-10 images of pr-vipe and ours in setting B. The query action is sit on the ground.

TABLE II
COMPARISON OF ACTION RETRIEVAL RESULT ON MOTION CAPTURE DATA WITH OCCLUSION.

	Recall (%) @ precision = 90%		
Experimental setting	A	B	C
pose [23]	17.8	63.2	11.8
pr-vipe [24]	43.0	14.7	9.7
ours	18.4	12.9	5.8
ours_lack	**78.2**	**77.8**	**48.6**

used for our feature to evaluate the retrieval performance with ignoring missing key-point (ours_lack) and with weighting characteristic key-points (ours_lack_weight) respectively. Given the matching distance, we use standard dynamic time warping (DTW) algorithm [40] to align two action sequences by minimizing the sum of frame matching distances, and sum of frame distances was used as the distance between two actions.

We reported the average precision and recall @ top-k with k = 1, 5, 10 on action retrieval in Table III and Table IV respectively. This shows that our proposed feature itself is less accurate by pr-vipe [24] method, but is able to detect similar action to query with higher precision and recall than the baseline methods with our flexible matching algorithm. Furthermore, since ours_weight_lack demonstrated the highest performance, retrieval with weighting characteristic key-points based on the proposed algorithm was effective.

TABLE III
COMPARISON OF ACTION RETRIEVAL PRECISION ON UT-KINECT DATASET [26].

	Average precision (%) @ top-k		
k	1	5	10
pose [23]	86.2	67.6	54.0
pr-vipe [24]	91.9	78.9	67.3
ours	91.7	75.2	63.4
ours_lack	95.5	85.3	78.3
ours_lack_weight	**96.0**	**86.3**	**79.6**

V. CONCLUSION

In this paper, we introduced our action retrieval technology realized by new view-invariant pose feature and flexible matching algorithm. The feature is calculated by 2D pose information, and has a one-to-one correspondence with each key-point of the original pose information. Matching algorithms can easily and flexibly ignore the missing key-points or weight characteristic key-points. Our method outperforms the state-of-the-art methods for pose retrieval tasks on an annotated 2D pose dataset and action retrieval task on a real-world video dataset.

TABLE IV
COMPARISON OF ACTION RETRIEVAL RECALL ON UT-KINECT DATASET [26].

	Average recall (%) @ top-k		
k	1	5	10
pose [23]	3.7	15.4	24.9
pr-vipe [24]	4.2	18.8	32.2
ours	4.2	17.7	30.0
ours_lack	4.6	20.7	38.3
ours_lack_weight	**4.6**	**21.0**	**39.0**

REFERENCES

[1] Tomi Räty. Survey on contemporary remote surveillance systems for public safety. *IEEE Trans. Syst. Man Cybern. Part C*, 40(5):493–515, 2010.

[2] Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):288–303, 2010.

[3] Sreemananth Sadanand and Jason J. Corso. Action bank: A high-level representation of activity in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 1234–1241. IEEE Computer Society, 2012.

[4] Yigithan Dedeoglu, B. Ugur Töreyin, Ugur Güdükbay, and A. Enis Çetin. Silhouette-based method for object classification and human action recognition in video. In Thomas S. Huang, Nicu Sebe, Michael S. Lew, Vladimir Pavlovic, Mathias Kölsch, Aphrodite Galata, and Branislav Kisacanin, editors, *Computer Vision in Human-Computer Interaction, ECCV 2006 Workshop on HCI, Graz, Austria, May 13, 2006, Proceedings*, volume 3979 of *Lecture Notes in Computer Science*, pages 64–77. Springer, 2006.

[5] Alireza Fathi and Greg Mori. Action recognition by learning mid-level motion features. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008.

[6] Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition from video using feature covariance matrices. *IEEE Trans. Image Process.*, 22(6):2479–2494, 2013.

[7] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. Action recognition by learning deep multi-granular spatio-temporal video representation. In John R. Kender, John R. Smith, Jiebo Luo, Susanne Boll, and Winston H. Hsu, editors, *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*, pages 159–166. ACM, 2016.

[8] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Spatiotemporal pyramid network for video action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2097–2106. IEEE Computer Society, 2017.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 568–576, 2014.

[10] Mahdyar Ravanbakhsh, Hossein Mousavi, Mohammad Rastegari, Vittorio Murino, and Larry S. Davis. Action recognition with image based CNN features. *CoRR*, abs/1512.03980, 2015.

[11] Dennis Ludl, Thomas Gulde, and Cristóbal Curio. Simple yet efficient real-time pose-based action recognition. In *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019, Auckland, New Zealand, October 27-30, 2019*, pages 581–588. IEEE, 2019.

[12] Federico Angelini, Zeyu Fu, Yang Long, Ling Shao, and Syed Mohsen Naqvi. Actionxpose: A novel 2d multi-view pose-based algorithm for real-time human action recognition. *CoRR*, abs/1810.12126, 2018.

[13] P V.V. Kishore, P Siva Kameswari, K Niharika, M Tanuja, M Bindu, D Anil Kumar, E Kiran Kumar, and M Teja Kiran. Spatial joint features for 3d human skeletal action recognition system using spatial graph kernels. *International Journal of Engineering & Technology*, 7(1.1):489–493, 2017.

[14] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7912–7921. Computer Vision Foundation / IEEE, 2019.

[15] Kalpit C. Thakkar and P. J. Narayanan. Part-based graph convolutional network for action recognition. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 270. BMVA Press, 2018.

[16] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Ahmed Sohel, and Farid Boussaïd. A new representation of skeleton sequences for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4570–4579. IEEE Computer Society, 2017.

[17] Eamonn J. Keogh, Themis Palpanas, Victor B. Zordan, Dimitrios Gunopulos, and Marc Cardle. Indexing large human-motion databases. In Mario A. Nascimento, M. Tamer Özsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer, editors, *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB 2004, Toronto, Canada, August 31 - September 3 2004*, pages 780–791. Morgan Kaufmann, 2004.

[18] X. Zhao, Myung Geol Choi, and Taku Komura. Character-object interaction retrieval using the interaction bisector surface. *Comput. Graph. Forum*, 36(2):119–129, 2017.

[19] Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph.*, 23(3):559–568, 2004.

[20] Jun Tang, Ling Shao, and Xiantong Zhen. Human action retrieval via efficient feature matching. In *10th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2013, Krakow, Poland, August 27-30, 2013*, pages 306–311. IEEE Computer Society, 2013.

[21] Arridhana Ciptadi, Matthew S. Goodwin, and James M. Rehg. Movement pattern histogram for action recognition and retrieval. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II*, volume 8690 of *Lecture Notes in Computer Science*, pages 695–710. Springer, 2014.

[22] Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. Pose embeddings: A deep architecture for learning to match human poses. *CoRR*, abs/1507.00302, 2015.

[23] Nataraj Jammalamadaka, Andrew Zisserman, Marcin Eichner, Vittorio Ferrari, and C. V. Jawahar. Video retrieval by mimicking poses. In Horace Ho-Shing Ip and Yong Rui, editors, *International Conference on Multimedia Retrieval, ICMR '12, Hong Kong, China, June 5-8, 2012*, page 34. ACM, 2012.

[24] Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 53–70. Springer, 2020.

[25] Stuart James, Manuel J. Fonseca, and John P. Collomosse. Reenact: Sketch based choreographic design from archival dance footage. In Mohan S. Kankanhalli, Stefan M. Rüger, R. Manmatha, Joemon M. Jose, and Keith van Rijsbergen, editors, *International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014*, page 313. ACM, 2014.

[26] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, pages 20–27. IEEE Computer Society, 2012.

[27] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7024–7033. Computer Vision Foundation / IEEE Computer Society, 2018.

[28] Girum G. Demisse, Konstantinos Papadopoulos, Djamila Aouada, and Björn E. Ottersten. Pose encoding for robust skeleton-based action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 188–194. Computer Vision Foundation / IEEE Computer Society, 2018.

[29] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4620–4628. Computer Vision Foundation / IEEE, 2019.

[30] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 53. BMVA Press, 2018.

[31] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3980–3989. IEEE Computer Society, 2017.

[32] Ankan Bansal. *Detecting and Recognizing Humans, Objects, and their Interactions*. PhD thesis, University of Maryland, College Park, MD, USA, 2020.

[33] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In Gang Hua and Hervé Jégou, editors, *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, volume 9914 of *Lecture Notes in Computer Science*, pages 627–642, 2016.

[34] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 472–487. Springer, 2018.

[35] Yadong Pan, Ryo Kawai, Noboru Yoshida, Hiroo Ikeda, and Shoji Nishimura. Training physical and geometrical mid-points for multi-person pose estimation and human detection under congestion and low resolution. *SN Comput. Sci.*, 1(4):208, 2020.

[36] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1302–1310. IEEE Computer Society, 2017.

[37] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2277–2287, 2017.

[38] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. Arttrack: Articulated multi-person tracking in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1293–1301. IEEE Computer Society, 2017.

[39] Georgia Gkioxari, Bharath Hariharan, Ross B. Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3582–3589. IEEE Computer Society, 2014.

[40] Petr Mandl and M. Rosario Romera Ayllón. On adaptive control of markov processes. *Kybernetika*, 23(2):89–103, 1987.