Continuous Action Recognition by Action-specific Motion Models

Keiji Morimoto, Yoichi Matsuyama, and Norimichi Ukita
Nara Institute of Science and Technology
e-mail ukita@is.naist.jp

Abstract

This paper proposes the models of human motion prior with multiple actions for action recognition in videos. A training sequence of each action, such as walking and jogging, is separately recorded by a motion capture system and modeled independently. Unlike existing approaches with similar motion prior models, our method uses the multiple models simultaneously for particle filtering in order to track the pose of a target person without being interfered by ambiguous motion. In addition to robustness to ambiguous motion, the particle transition among multiple motion models allows us to continuously identify the action of the target person frame-by-frame.

1 Introduction

This paper proposes continuous action recognition in videos. Our proposed method achieves frame-by-frame action recognition by continuous pose tracking, while previously ones classifies a sequence.

For accurate and robust pose tracking, motion prior[1, 2, 3] is effective. The precise prior of a human body can be obtained by a motion capture system. Many kinds of actions, such as walking and running, have been recorded in datasets[4, 5] that are widely used for modeling and evaluating human motion. The motion model of each action (i.e. action-specific motion model) can be used for pose tracking in that action. Finding the proper action model at each moment corresponds to action recognition. That is the proposed scheme for action recognition.

However, the datasets described above have only elemental actions (e.g. walking and jogging) but no transitions among the actions (e.g. from walking to jogging). Since potential transitions among all of the actions are extremely varied (i.e. intra-individual and inter-individual variations), recording all of these variations is unrealistic. To cope with this problem, in the proposed method, smooth particle transitions among the elemental actions are explicitly represented by synthesized transition paths. With the particle transitions among action-specific models, the goal of this work is to achieve frame-by-frame action recognition.

2 Related Work

While particle filtering[6] is effective, for tracking robust to failure in image processing, it is not easy to apply it to human pose tracking due to two reasons: 1) curse of dimensionality and 2) complex body-motion. Since a human body is high-dimensional data, 1) it is difficult to distribute particles sufficiently in the high-dimensional space and 2) generalized motion prediction for particle transition is difficult due to inter-individual variation in motion.

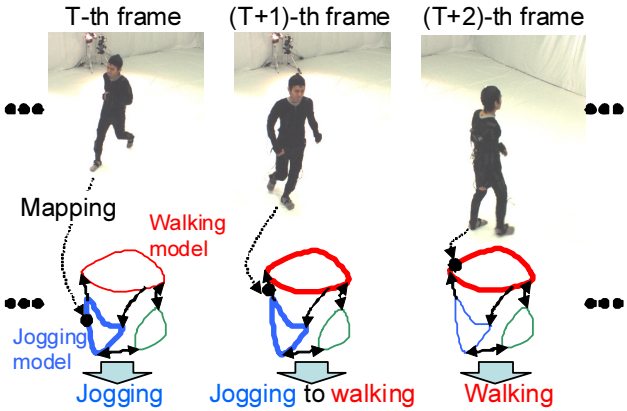

Figure 1. Continuous frame-by-frame action recognition. A feature extracted from an observed image is mapped to action-specific motion models (indicated by red, blue, and green closed curves). A mapping point (indicated by black points) is acquired robustly by particle filtering in the models. The model having the maximum affinity for the mapping point is selected and then its respective action is regarded as an action observed at each frame.

In most of motion models, high-dimensional motion data is modeled in lower dimensions in order to resolve the problems mentioned above. For low-dimensional modeling, nonlinear probabilistic embedding such as Gaussian Process Latent Variable Models (GPLVM[7]) is widely used. Several extensions of GPLVM have been studied, for example, for modeling motion dynamics (Gaussian Process Dynamical Models, **GPDM**, in short)[1] and dependency between different kinds of data that have the similar structures[8].

The latent models allow us to model multiple kinds of actions as well as a single action; a set of independently trained models[2, 3] and a unified model with multiple actions trained together[2, 9, 10]. While selection of an appropriate model at each moment is required for a set of the independent models, it has several advantages; 1) since the computational cost of modeling grows as sample data increase in each model, the independent models can be computed fast and 2) each model is optimized for its respective action.

Synthesizing realistic transitions between different poses has been studied in Computer Graphics, such as **motion graphs**[11].

3 Pose Tracking in a Single Action

Our pose tracking is based on pose regression from image features. This section describes its previous methods for a single action.

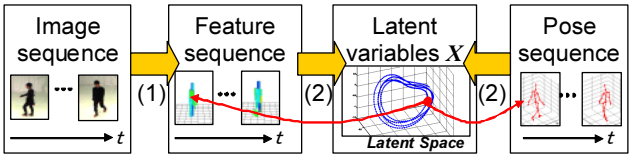

Figure 2. Learning motion prior and feature-to-pose regression of a single action. (1) Feature extraction (e.g. shape contexts[14]). (2) Shared latent-structure modeling with GPDM. Each latent variable corresponds to its respective feature and pose as depicted by red arrows.

3.1 Motion Modeling by GPDM

Gaussian Process Dynamical Models (**GPDM**)[1] acquire smooth dynamics of sample data in a low-dimensional latent space X from high-dimensional observation data, joint positions in our experiments. GPDM is defined by two mapping functions; 1) from a point at t to a point at $t+1$ in the latent space, $f_D(\boldsymbol{x})$ where $\boldsymbol{x} \in X$, and 2) from the latent space to the observation space, $f_O(\boldsymbol{x})$. $f_D(\boldsymbol{x})$ gives us the capability of prediction that is useful for tracking (e.g. used for particle filtering[12]). The nature of GP also provides the distribution (variance) of data.

Given a sample sequence with N frames $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N]$, the mapping functions are acquired by maximizing the joint likelihood of $\boldsymbol{Y}$ and $\boldsymbol{X}_{t+1}$ with respect to $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ and $\boldsymbol{X}_t$, respectively, where $\boldsymbol{X}_{t+1} = [\boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$ and $\boldsymbol{X}_t = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1}]$. In this optimization, similarity between components $\boldsymbol{x}_i$ in $\boldsymbol{X}$ is evaluated by RBF.

3.2 Mapping between Feature and Pose Spaces by a Shared Latent Structure

For feature-to-pose regression, synchronized features and poses ("Feature sequence" and "Pose sequence" in Fig. 2) are modeled together. The connection between them is established by the latent space shared by their observation spaces, as shown by (2) in Fig. 2. With the shared structure, each latent variable $\boldsymbol{x}_i$ corresponds to its respective pose and feature as depicted by thin red arrows in Fig. 2

3.3 Particle Tracking for Pose Regression

Feature tracking is performed for feature-to-pose regression at each moment. Feature tracking is achieved by particle filtering in X. Each particle in X corresponds to a feature. At each frame t, a feature is extracted from a captured image. The particles transit from the ones at the previous frame (frame $t-1$) using motion prior $f_D(\cdot)$ and then are mapped to the feature space. The likelihood $c_{t,i}$ of i-th particle at t (denoted by $\boldsymbol{x}^p_{t,i}$) with regard to the feature at t (denoted by $\boldsymbol{f}_t$) is expressed as follows:

$$c_{t,i} = \exp\left(-w_v \sigma^2_{t,i} - w_o ||f^F_O(\boldsymbol{x}^p_{t,i}) - \boldsymbol{f}_t||^2\right), \qquad (1)$$

where $\sigma^2_{t,i}$ and $f^F_O(\cdot)$ denote the variance of $\boldsymbol{x}^p_{t,i}$ in the model and the mapping function from X to the feature space, respectively. Weight variables w_v and w_o are given empirically; $w_v = 0.5$ and $w_o = 0.5$.

Finally, the pose is estimated by mapping the likelihood-weighted mean of the particles from X to the pose space.

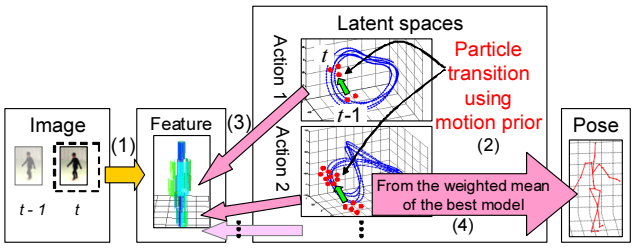

Figure 3. Pose tracking with multiple motion models. (1) Feature extraction. (2) Particle tracking using motion prior. (3) Mapping the particles into the feature space for evaluating their likelihood. (4) Mapping the likelihood-weighted sum of the particles from the latent space that has the maximum likelihood into the pose space.

4 Continuous Action Recognition by Pose Tracking using Action-specific Motion Models

4.1 Particle Tracking by Simultaneously using Multiple Action-specific Models

In pose tracking with multiple action models (Fig. 3), the model corresponding to an action observed at each moment should be selected for correct pose tracking. Existing tracking methods with multiple motion models[3] have the problems of unrobust tracking due to using a single model at each moment and no transition path among the models.

To resolve these two problems, the following ideas are employed in our proposed method:

- **Multiple hypotheses of motion:** Particles are distributed in multiple models simultaneously for multiple hypotheses of motion prior.
- **Synthesized transition path:** Transition paths are synthesized from the real samples of multiple actions. Motion dynamics along the synthesized paths leads the particles smoothly to a next action model.

Pose tracking with our motion models is designed in accordance with Condensation[6]:

1. Particles are distributed in all models.
2. The particles are drifted using motion prior $f_D(\cdot)$ and diffused at t so that more particles are placed near the ones having higher likelihood at $t-1$, as illustrated in (2) in Fig. 3. Note that the particles are distributed simultaneously in multiple models.
3. Each particle is mapped to the feature space and compared with the feature at t for evaluating the likelihood of the particle by Eq. (1), as depicted by (3) in Fig. 3.

4. The total sum of the likelihoods of all particles in m-th model is considered to be the goodness-of-fit between the model and the feature observed at t. The goodness-of-fit, $F_{m,t}$, is expressed as follows:

$$F_{m,t} = \sum_i c_{t,i}^m, \quad (2)$$

where $c_{t,i}^m$ denotes The model having the maximum goodness-of-fit is selected. This model is denoted by M^{max}. The likelihood-weighted sum of the particles in the selected model, M^{max}, is mapped to the pose space for estimating the pose observed at t, as shown by (4) in Fig. 3.

Diffusion among models is achieved via transition paths, as well as within a model. Synthesizing the transition paths is described in Sec. 4.2, followed by inter-model diffusion via them described in Sec. 4.3.

4.2 Synthesizing Transition Paths among Multiple Actions

As with motion graphs[11], the end points of each transition path are determined so that their respective poses have the local maximum of similarity between pose vectors $\boldsymbol{y}$ of two actions a and b. The similarity is expressed by $-||\boldsymbol{y}_i^a - \boldsymbol{y}_j^b||$, where subscripts denote i-th and j-th frames. $\boldsymbol{y}$ consists of 3D positions of all joints.

New paths are synthesized by interpolating sample poses. Compared with naive interpolation[11], good connectivity can be achieved by finding the shortest path between the samples via a number of intermediately generated interpolated poses in [13].

In the proposed model, an additional constraint is employed by taking into account smoothness in the sample motion. The existing methods[11, 13] control smoothness of the transition by adjusting the number of interpolating points. How to determine the number of the points is important, which has not been discussed in [11, 13] Our method determines the number of the points so that the curvature of the synthesized path is less than the largest one in the sample motion. Specifically, 1) interpolating points are initially located at regular intervals, each of whose length is equal to the maximum length between successive sample points (denoted by l^{max}), between a pair of end points and then 2) the end points are shifted along their successive sample points until all angles between successive interpolating points are less than the maximum angle in sample points, while the interpolating points are increased so that the length between them is less than l^{max}.

4.3 Particle Propagation via Transition Paths

Unlike Condensation[6], the proposed method distributes particles in multiple models and propagates them among the models as follows.

In each model, all particles are diffused after drift using motion prior. Diffusion is represented by the following equation:

$$p(\boldsymbol{x}^d|\boldsymbol{x}^m) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}^d - \boldsymbol{x}^m)^2\right), \quad (3)$$

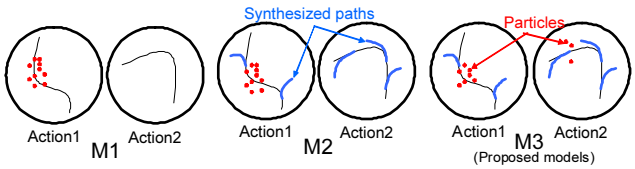

Figure 4. Motion models and particles. Solid black and blue lines depict real sample and synthesized data. Red dots depict particles.

where $\boldsymbol{x}^m$ and $\boldsymbol{x}^d$ denote particles given by drift using motion prior and diffusion, respectively. Particles each of whose nearest sample is one of the synthesized samples are selected. Let m and n denote the models where the selected particle is located and the one to which m connects via the synthesized path, respectively. For each of the selected particles, the variances ${\sigma_m}^2$ in m and ${\sigma_n}^2$ in n are computed. Then there is $\exp(-{\sigma_n}^2)$ in $(\exp(-{\sigma_m}^2) + \exp(-{\sigma_n}^2))$ chance that the particle moves to model n.

4.4 Action Recognition by Particles

An action observed at each moment is classified based on the distribution of particles. The simplest way is that the model M^{max} having the maximum of the goodness-of-fit, each of which is expressed by Eq. (2), is selected as the one corresponding to the action observed at each frame. However, this naive selection might fluctuate the result of action recognition, in particular immediately before, immediately after, and during action transitions. As well as such transitional states, temporarily-similar poses in different actions also cause the fluctuation of the recognition results.

To cope with this problem, the goodness of fit in each model is smoothed by the Kalman filter. While the Kalman filter cannot represent quick and sharp changes, this disadvantage is acceptable because human action is not changed so quickly. By selecting the model having the maximum response of the Kalman filter, frame-by-frame action recognition is achieved.

5 Experiments

5.1 Image Features and Datasets

The bag-of-shape-contexts[14] were used for empirical evaluation. A feature at each frame was represented by a set of the shape contexts obtained from the silhouette of a target person; 200 points sampled along the boundary of the silhouette. Dissimilarity between two features, $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, is expressed by $||\boldsymbol{f}_1 - \boldsymbol{f}_2||^2$.

Synchronized video and mocap datasets of multiple actions were used for learning and evaluation. The videos were captured at 30 fps (1024×768 pixels). For learning and ground-truth, a gyro-sensor based motion capture system (IGS-190) was used. 51 dimensional pose data (i.e. 17 3-DOF joints) was obtained.

Three kinds of action sets below were used:

- **Set1 (four dance actions):** Waving the arms by different four ways.

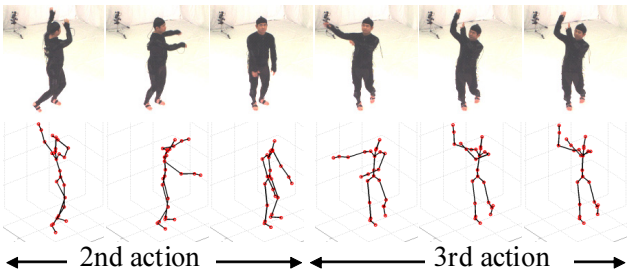

Figure 5. Action recognition results in action set1. Top: observed images. Middle: Estimated poses. Bottom: Estimated actions.

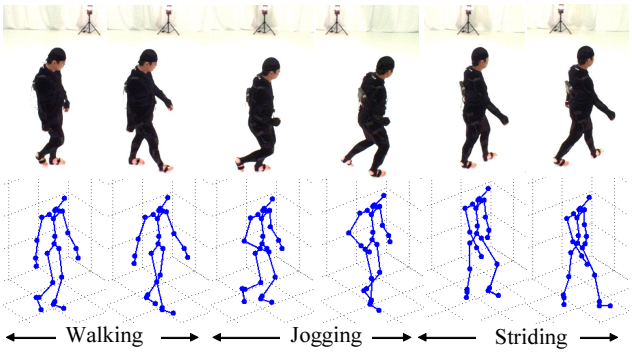

Figure 6. Action recognition results in action set3.

Table 1. Percentages of correctly classified actions at each frame.

	set1	set2	set3
M1 (no transition paths)	72.3	84.5	70.1
M2 (single motion prior)	78.2	85.9	71.8
M3 (proposed)	80.6	88.1	75.7

- **Set2 (two gait actions):** Walking and jogging.
- **Set3 (six gait actions):** 1) Walking, 2) walking slowly, 3) walking fast, 4) striding, 5) jogging, and 6) stopping from walking and start walking.

Our datasets contain a number of transitions between each pair of actions, which are required for validating the proposed models. The type of action at each frame in test sequences was labeled manually for evaluation. While only one subject was captured for learning samples, five subjects were captured for evaluation.

With each of the action sets, three kinds of models were tested (Fig. 4): M1) with no synthesized paths, M2) with synthesized paths but using unimodal motion prior at each moment, where all particles propagated in a single model at each moment, and M3) with synthesized paths using motion prior of multiple actions models (proposed models).

5.2 Recognition Results

Figures 5 and 6 show the results of pose tracking and correctly-estimated actions in set1 and set3, respectively. While the difference among the estimated poses of different actions is small, the actions could be classified correctly.

Table 1 shows recognition rates per frame. In all datasets, set1, set2, and set3, the proposed model could outperform other models.

6 Concluding Remarks

This paper proposed the prior models of multiple actions for continuous action recognition by pose tracking. The models are acquired from independently captured action sequences so that potential transition paths between them are synthesized. Experimental results demonstrated that the synthesized paths and particles propagated in multiple models allow us to more correctly classify observed actions frame-by-frame.

References

[1] J. M. Wang, D. J. Fleet, A. Hertzmann, "Gaussian Process Dynamical Models for Human Motion," *PAMI*, Vol.30, No.2, pp.283–298, 2008.

[2] R. Urtasun, D. J. Fleet, A. Geiger, J. Popovic, T. Darrell, and N. D. Lawrence, "Topologically-Constrained Latent Variable Models," *ICML*, 2008.

[3] J. Chen, M. Kim, Y. Wang, and Q. Ji, "Switching Gaussian Process Dynamic Models for Simultaneous Composite Motion Tracking and Recognition," *CVPR*, 2009.

[4] L. Sigal, A. Balan and M. J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *IJCV*, Vol.87, No.1–2, 2010.

[5] CMU Graphics Lab Motion Capture Database: http://mocap.cs.cmu.edu/

[6] M. Isard and A. Blake, "CONDENSATION - Conditional Density Propagation for Visual Tracking," *IJCV*, Vol.29, No.1, pp.5–28, 1998.

[7] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, Vol.6, pp.1783–1816, 2005.

[8] A. P. Shon, K. Grochow, A. Hertzmann, and R. P. N. Rao, "Learning Shared Latent Structure for Image Synthesis and Robotic Imitation," *NIPS*, 2005.

[9] A. Geiger, R. Urtasun, and T. Darrell, "Rank Priors for Continuous Non-Linear Dimensionality Reduction," *CVPR*, 2009.

[10] L. Raskin, M. Rudzsky, and E. Rivlin, "Dimensionality reduction using a Gaussian Process Annealed Particle Filter for tracking and classification of articulated body motions," *CVIU*, Vol.115, Issue 4, pp.503–519, 2011.

[11] L. Kovar, M. Gleicher, and F. H. Pighin, "Motion graphs," *SIGGRAPH*, 2002.

[12] N. Ukita, M. Hirai, and M. Kidode, "Complex Volume and Pose Tracking with Probabilistic Dynamical Models and Visual Hull Constraints," *ICCV*, 2009.

[13] L. Zhao and A. Safonova, "Achieving good connectivity in motion graphs," *Graphical Models Journal*, Vol.71, No.4, pp.139–152, 2009.

[14] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *PAMI*, Vol.24, No.4, pp.509–522, 2002.