SCIENCE CHINA
Information Sciences

• LETTER •

March 2022, Vol. 65 139101:1–139101:3
https://doi.org/10.1007/s11432-019-2869-x

Self-adjustable hyper-graphs for video pose estimation based on spatial-temporal subspace construction

Jizhou MA[1], Shuai LI[1*], Hong QIN[2], Aimin HAO[1] & Qinping ZHAO[1]

[1]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing* 100191*, China;*
[2]*Department of Computer Science, Stony Brook University, New York* 11794*, USA*

Received 7 October 2019/Revised 25 January 2020/Accepted 7 April 2020/Published online 20 May 2021

Citation Ma J Z, Li S, Qin H, et al. Self-adjustable hyper-graphs for video pose estimation based on spatial-temporal subspace construction. Sci China Inf Sci, 2022, 65(3): 139101, https://doi.org/10.1007/s11432-019-2869-x

Dear editor,

In recent years many supervised video pose estimation methods have achieved growing successes based on well-labeled training datasets. Nonetheless, when facing roughly-labeled training data, it still remains challenging to intrinsically encode the video contents' spatial-temporal coherency for robust video pose estimation.

Some researches aimed to directly improve and refine the existing confidence maps by combining the spatial-temporal structure models [1, 2]. Li et al. [2] suggested that fixing some reliable estimations and formulating propagation processing as a 3D trajectory completion problem. Differently, Moon et al. [1] assumed that state-of-the-art 2D human pose estimation methods have similar error distributions. Zhou and Torre [3] first learned a codebook from the motion capture dataset, and then they employed a bi-linear model to estimate poses by matching the movement mode and the dense trajectory tracing result. It enables related patterns to be expressed using sub-patterns instead of a uniform probability distribution. In [3], this flexible codebook based framework still requires a lot of extra annotated data to ensure the models' dataset-specific applicability.

To overcome this drawback, we advocate a new hierarchical hyper-graph approach based on intrinsic spatial-temporal subspace exploration and propagation. Those "mis-matched" hyper-graph subspaces, which result from imperfect data, could be adaptively improved by taking advantage of visual contents' intrinsic continuities. At the theoretic level, the key idea for subspace exploration is to design a maximum matching subspace (MMS) operator, which help propagate highly correlated action information from local video frames to all video sequences in spatial-temporal subspaces. The hyper-graph is solely built based on our MMS metric, and it could synchronously encode cross-video action similarity, inner-video temporal coherency, and synergetic relationship of different body joints.

In contrast to normal "explicit hyper-graph", we construct an "implicit hyper-graph" by hierarchically representing different-level relationships. We conceptually split "explicit hyper-graph" into a series of sub-graphs (structure), which are formulated as optimized maximum matching subspaces. Then, these subspaces ("sub-graphs") will re-contact with each other via a global MMS operator based affinity matrix (metric).

Given a set of videos belonging to the same action category, each video is divided into a group of overlapping short video segments. The initial pose extractor ResNet50 [4], and such segments are represented as N_P pose sequences $\boldsymbol{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \ldots, \boldsymbol{P}_{N_P}\}$. For a pose sequence $\boldsymbol{P}_i \in \mathbb{R}^{n_k \times n_f}$, it covers n_k body joints and n_f consecutive frames. To align two pose sequences, similar to [3], we apply Procrustes analysis to get a spatial transition matrix $\boldsymbol{Q}$, which conducts an affine transformation, including translating, rotating and uniformly scaling. We mark it as $\boldsymbol{Q}_i(\boldsymbol{P}_j) : \boldsymbol{P}_i \overset{\boldsymbol{Q}_i}{\leftarrow} \boldsymbol{P}_j$, which means the pose sequence $\boldsymbol{P}_j$ is aligned to $\boldsymbol{P}_i$ by the transformation $\boldsymbol{Q}_i$, and $\boldsymbol{Q}_i(\boldsymbol{P}_j)$ represents the transformed matrix.

$$\alpha_{i,j} = \arg\min_{\alpha_{i,j}} ||\boldsymbol{P}_i - \alpha_{i,j}\boldsymbol{P}_j||_{\mathrm{F}}. \tag{1}$$

Here $\alpha_{i,j}$ is the scaling coefficient, and it is the only parameter that needs to be optimized. $\alpha_{i,j}$ may not equal $\alpha_{j,i}$, and thus $||\boldsymbol{P}_i - \alpha_{i,j}\boldsymbol{P}_j||_{\mathrm{F}}$ is not equal to $||\boldsymbol{P}_j - \alpha_{j,i}\boldsymbol{P}_i||_{\mathrm{F}}$ as well. The transformed affinities of the two pose sequences are asymmetric.

MMS. Most methods measure pose sequence affinities with a simple Frobenius-norm $||\boldsymbol{P}_i - \boldsymbol{P}_j||_{\mathrm{F}}$, which misses a synergetic relationship analysis over joints and frames. It means that, two adjacent pose sequences in a video may be greatly different in the pose space because of temporal topology dislocation.

$$\psi_M(\boldsymbol{P}_i, \boldsymbol{P}_j \,|\, \boldsymbol{\Theta}) = ||\boldsymbol{P}_i(K, F_i) - \boldsymbol{Q}_i(\boldsymbol{P}_j)(K, F_j)||_{\mathrm{F}}. \tag{2}$$

Here $\boldsymbol{\Theta} = \{K, F_i, F_j\}$, (K, F_i), and (K, F_j) are the subspace parameters, which actually is the sub-matrix index and is determined by the specific sequence pair. $\{K, F_i, F_j\}$ respectively denotes the set (indexes) of joints, frame in $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$.

* Corresponding author (email: lishuai@buaa.edu.cn)

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2021

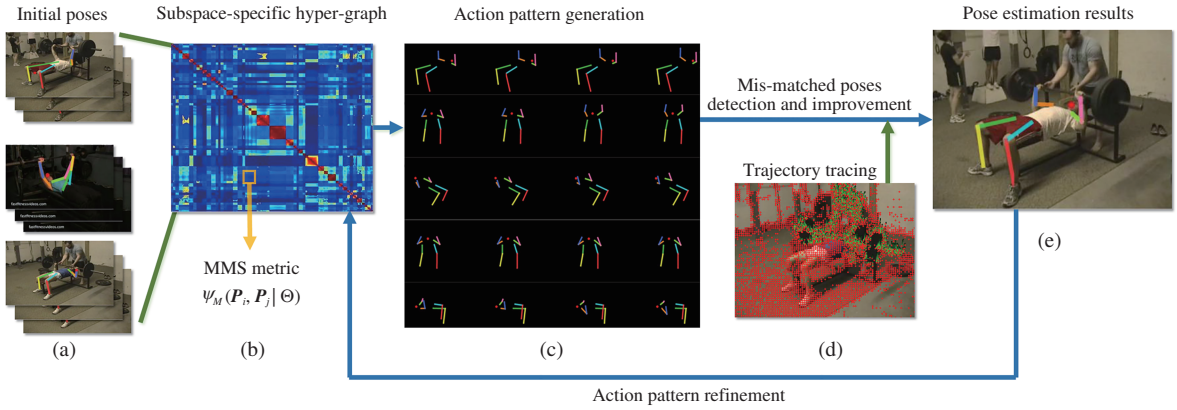

Figure 1 (Color online) The pipeline of our framework. (a) Input video sequences within the common action category; (b) the MMS metric governed subspace-specific hyper-graph construction; (c) generating the action pattern which consists of a series of clustered action sub-patterns; (d) the dense trajectory tracing for local details capturing; (e) detecting and improving existing mis-matched (inaccurate) poses by incorporating the semantic guidance from the action patterns and local feature tracing from the dense trajectories.

Both of the two sequences have the same number of frames ($|F_i| = |F_j|$). As a result, ψ_M partitions the difference of $\boldsymbol{P}_i$ and $\boldsymbol{P}_j$ into aligned matching sub-matrix $\boldsymbol{\Theta}$ and unmatched sub-matrix. In this study, f/F indicates frames, k/K indicates body joints, and $\boldsymbol{\Theta}$ represents both of them. The $\boldsymbol{\Theta}$ can be efficiently computed via a simple path searching algorithm. The searching space encodes spatial-temporal affinity in a 3-dimension matrix $\boldsymbol{D}_{i,j} \in \mathbb{R}^{n_f \times n_f \times n_k}$. Its element is defined as

$$D(k, f_i, f_j) = \boldsymbol{D}_{k,f_i,f_j} = ||\boldsymbol{P}_i(k, f_i) - \boldsymbol{Q}_i(\boldsymbol{P}_j)(k, f_j)||_2. \quad (3)$$

Then, the path is partitioned at the steepest point with the maximum gradient. The first half of the path indexes the maximum matching subspaces. On the other hand, the rest of the subspaces denotes those dissimilar parts of pose sequences.

Action pattern generation. The action expression encoded in our hyper-graph is comprehensive but losing generality. Thus, we explore some mid-level semantic sub-patterns (sub-actions) to describe a complete action. These sub-patterns can be considered as certain words in "action codebook" to describe the pose sequences.

To begin with, we build an affinity matrix $\boldsymbol{A} \in \mathbb{R}^{N_P \times N_P}$ for all the input pose sequences, with

$$\begin{aligned}\boldsymbol{A}(i,j) &= \boldsymbol{A}(j,i) \\ &= \frac{1}{2}\psi_M(\boldsymbol{P}_i, \boldsymbol{Q}_i(\boldsymbol{P}_j)|\boldsymbol{\Theta}) + \frac{1}{2}\psi_M(\boldsymbol{P}_j, \boldsymbol{Q}_j(\boldsymbol{P}_i)|\boldsymbol{\Theta}). \quad (4)\end{aligned}$$

Based on $\boldsymbol{A}$, we cluster all the pose sequences into N_B groups via spectral clustering. For the c-th group, we conduct hierarchical clustering and orderly merge the pose sequences one by one to construct a sub-pattern $\boldsymbol{B}_c$. Yet, in the following step, we also need to trade off the contribution of those densely sampled trajectories [5] for the local detail tracing. Therefore, we adopt the 2D Gaussian function to define motional joint models.

As for "mean" pattern (pose sequence) $\boldsymbol{\mu} \in \mathbb{R}^{n_k \times n_f}$, it has the same formulation as a pose sequence $\boldsymbol{P}_i$. $\boldsymbol{B}_c$ is built by aggregation processing over n_c pose sequences $\boldsymbol{P}_n$ included in $\boldsymbol{B}_c$, as

$$\begin{aligned}&\boldsymbol{B}_c^{(n+1)}(\boldsymbol{\Theta}) = \boldsymbol{B}_c^{(n)}(\boldsymbol{\Theta}) + \lambda^{(n)} \cdot \boldsymbol{P}_n(\boldsymbol{\Theta}),\ n \in [1, n_c] \\ &\text{s.t.}\ \ \boldsymbol{B}_c^{(0)} = \boldsymbol{0}^{n_k \times n_f} \text{ and } \boldsymbol{\Theta} \in \psi_M(\boldsymbol{B}_c^{(n)}, \boldsymbol{P}_n \mid \boldsymbol{\Theta}).\end{aligned} \quad (5)$$

Here $\lambda^{(n)}$ denotes the hierarchical clustering distance from sub-pattern $\boldsymbol{B}_c$ to pose sequence $\boldsymbol{P}_n$. $\boldsymbol{\Theta}$ is computed by MMS operator, which changes dynamically corresponding to current $\boldsymbol{B}_c^{(n)}$ and $\boldsymbol{P}_n$. After constructing all sub-patterns $\boldsymbol{B}_c$, we can obtain a complete action pattern as $\boldsymbol{B} = \{\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_{N_B}\}$.

We have defined our MMS operator $\psi_M(\boldsymbol{P}_i, \boldsymbol{P}_j \mid \boldsymbol{\Theta})$ for two pose sequences in Eq. (2), and we also need to use it on action sub-patterns as $\psi_M(\boldsymbol{P}_i, \boldsymbol{B}_c \mid \boldsymbol{\Theta})$. Therefore, we redefine the distance matrix in (3) for adapting MMS to the probabilistic sub-patterns $\boldsymbol{B}_c = \{\boldsymbol{P}_\mu, \rho\}$, with

$$\begin{aligned}&\boldsymbol{d} = ||\boldsymbol{P}_i(k, f_i) - \boldsymbol{Q}_i(\boldsymbol{B}_c)(k, f_c)||_1, \\ &D_{\text{prob}}(\boldsymbol{d} \mid \boldsymbol{B}_c) = \boldsymbol{D}_{k,i,c} = \frac{1}{2\pi|\rho|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{d}^{\mathrm{T}}\rho^{-1}\boldsymbol{d})\right),\end{aligned} \quad (6)$$

where $\boldsymbol{d} \in \mathbb{R}^{2\times 1}$ denotes the distance of joint $\boldsymbol{P}_i(k, f_i)$ to the mean joint position of sub-patterns $\boldsymbol{B}_c(k, f_c)$. $\boldsymbol{D}_{i,c}$ shows the 2D Gaussian function based affinity. In Eq. (3), $\boldsymbol{D}$ is a distance (dissimilarity) matrix, but in (6), $\boldsymbol{D}$ shows the probabilistic affinity (similarity).

Considering the confused spatial-temporal continuity over initial poses, frame-wisely refining the pose is more robust than global sequence-wise processing. The f-th frame error $E(\boldsymbol{p}_f)$ is gotten by averaging the reconstruction error overall pose sequences covering the f-th frame ($\boldsymbol{p}_f \in \boldsymbol{P}_i$). The error E of f-th frame's pose $\boldsymbol{p}_f$ is defined as

$$E(\boldsymbol{p}_f) = E_{\text{tra}}(\boldsymbol{p}_f \mid \dot{\boldsymbol{B}}_c) + \beta \cdot E_{\text{act}}(\boldsymbol{p}_f \mid \dot{\boldsymbol{B}}_c), \quad (7)$$

where E_{tra} denotes the trajectory tracing error and E_{act} means the action pattern governed reconstruction error. The matched sub-pattern $\dot{\boldsymbol{B}}_c$ guides the computation procedure of E_{tra} and E_{act}. To compute E_{tra}, $\dot{\boldsymbol{B}}_c$ determine the weight of each trajectory around the joints. As for E_{act}, $\dot{\boldsymbol{B}}_c$ measures the action fitting error. We find the best matched sub-pattern for each pose sequence, with

$$(\boldsymbol{P}_i, \dot{\boldsymbol{B}}_c) = \arg\min_{\boldsymbol{P}_i, \boldsymbol{B}_c} \psi_M(\boldsymbol{P}_i, \boldsymbol{B}_c \mid \boldsymbol{\Theta}), \quad (8)$$

where $\boldsymbol{B}_c \in \boldsymbol{B}$ and all the related pose sequences $\boldsymbol{P}_i$ should cover the f-th frame, $\boldsymbol{P}_i(\boldsymbol{\Theta}) \ni \boldsymbol{p}_f$. $\boldsymbol{\Theta}$ indexes a sub-matrix of $\boldsymbol{P}_i$, and $\boldsymbol{P}_i$ may cover $\boldsymbol{p}_f$ but $\boldsymbol{P}_i(\boldsymbol{\Theta})$ not.

$\mathbb{T}$ is a local feature tracing function, and it gives the trajectory tracing results $\boldsymbol{t} \in \mathbb{R}^{n_k \times 1}$ related to $\boldsymbol{p}_f$ from the last $(f-1)$-th frame $\boldsymbol{t}^{(0)}$ to the current f-th frame $\boldsymbol{t}^{(1)}$. E_{tra} is formulated as

$$E_{\text{tra}}(\boldsymbol{p}_f) = \frac{1}{|\mathbb{T}(\boldsymbol{p}_f)|} \cdot \sum_{\boldsymbol{t} \in \mathbb{T}(\boldsymbol{p}_f)} ||D_{\text{prob}}(\boldsymbol{t}^{(1)} \mid \dot{\boldsymbol{B}}_c) - D_{\text{prob}}(\boldsymbol{t}^{(0)} \mid \dot{\boldsymbol{B}}_c)||_2, \tag{9}$$

where E_{tra} is able to detect the errors caused by drastic pose location changes. Moreover, we define the action pattern based reconstruction error as

$$E_{\text{act}}(\boldsymbol{p}_f) = \frac{1}{n_m} \sum_{\boldsymbol{P}_i \ni \boldsymbol{p}_f} \psi_M(\boldsymbol{P}_i, \dot{\boldsymbol{B}}_c \mid \boldsymbol{\Theta}(f)). \tag{10}$$

Here $\boldsymbol{\Theta}(f)$ denotes the frame-specific parameter, which only indexes the matching joints set K at the f-th frame. n_m is the size of the candidate matching set. In practice, we do not use all eligible $\boldsymbol{P}_i$ and only pick 5 ($n_m = 5$) pose sequences with the minimum matching errors (Eq. (8)).

Finally, as shown in Figure 1, we formulate an iterative pose estimation framework, which alternately conducts pose improvement and action pattern refinement. The intermediately improved poses facilitate pattern amending. We fix the top 5% frames (sorted by their errors $E(\boldsymbol{p}_f)$) for pose improvement in each loop.

Conclusion. We have detailed a spatial-temporal subspace involved hyper-graph method for human pose estimation in video analysis. The method can indeed improve the existing estimated results even without the need for labeled ground-truth poses. All the exhibited advantages of the new method result from the MMS operator, which enables intrinsic encoding of the action similarity between videos, the intra-video temporal coherency, and the collaborative relevance over different body joints.

Acknowledgements This work was supported in part by National Key R&D Program of China (Grant No. 2018YFB-1700603), National Natural Science Foundation of China (Grant Nos. 61672077, 61532002), and Beijing Natural Science Foundation — Haidian Primitive Innovation Joint Fund (Grant No. L182016).

References

1 Moon G, Chang J Y, Lee K M. Posefix: model-agnostic general human pose refinement network. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019
2 Li Z, Wang X, Wang F, et al. On boosting single-frame 3D human pose estimation via monocular videos. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019
3 Zhou F, de la Torre F. Spatio-temporal matching for human pose estimation in video. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 1492–1504
4 Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking. In: Proceedings of European Conference on Computer Vision, 2018
5 Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of IEEE International Conference on Computer Vision, 2013