Neurocomputing 166 (2015) 301–308

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Human pose recovery by supervised spectral embedding

Jun Yu [a], Yukun Guo [b], Dapeng Tao [c,*], Jian Wan [a]

[a] Hangzhou Dianzi University, China
[b] Xiamen University, China
[c] School of Information Science and Engineering, Yunnan University, China

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form
17 March 2015
Accepted 3 April 2015
Communicated by M. Wang
Available online 23 April 2015

Keywords:
Human pose estimation
Spectral embedding
k-NN regression

ABSTRACT

In this paper we propose a subspace learning algorithm based on supervised manifold learning techniques to address the problem of inferring 3D human poses from monocular video frames. Low-dimensional representations of visual features are computed via spectral embedding, regularized by the pairwise relationship of poses for simultaneously preserving the locality in the feature space and taking account of similarities in the pose space. To deal with the "out-of-sample" problem, we obtain a global linear projection from the embedding whereby the Euclidean distances between transformed feature vectors can faithfully reflect the corresponding pose distances. To retrieve the most similar candidate from the exemplar database, weighted sum of Euclidean distances of features is employed to achieve better accuracy instead of simply summing up the squared distances of all feature types. The experimental results on HumanEva dataset validate the efficacy of our proposed method.

1. Introduction

Vision based human pose estimation is important for a wide range of applications, including video surveillance, human–computer interaction (HCI) and semantic annotation of video archives. The inherent challenges such as the high dimensionality of the pose space to search and great variation of human appearances, make human pose estimation remain a largely unsolved problem in spite of many years of research.

In this paper, we consider the task of recovering 3D human poses from a static monocular camera. The research in this field can roughly be classified into two main categories: the generative (model-based) approaches and the discriminative (model-free) ones. The discriminative approaches directly model the mapping from visual inputs to the high degree-of-freedom body configuration as a multivariate regression problem [1]. Or more straightforwardly, this mapping can be explicitly obtained via exemplar-based learning [2]. Provided an exemplar database endowed with visual features and the corresponding poses, which is sufficiently dense and versatile to cover the variations of human postures, exemplar-based approaches can easily recover the pose from a novel video frame by retrieving the most visually similar instance in the database using nearest neighbor search.

The accuracy of the recovered poses substantially depends on the image features. It is common practice to extract multiple types of features for better description of the visual input. However, simply concatenating multi-view features into a high-dimensional vector suffers from the "curse of dimensionality" problem, which degrades the performance of the estimation system. To this end, applying dimension reduction/subspace learning methods prior to *k*-NN algorithm is crucial. Subspace learning aims to find the optimal projection matrix whereby the Euclidean distance in the projected low-dimensional subspace can capture the intrinsic relationship of data points. As a special case, feature selection finds the most efficient feature subset by keeping useful feature components while discarding irrelevant or misleading ones [3,4]. In feature selection, the projection matrix is constrained to a selection matrix, and it can improve both accuracy and efficiency. By contrast, general subspace learning has the potential to achieve better accuracy with such constraint eliminated. The obvious shortcoming of subspace learning is the high computational complexity of the training stage and the need to perform projection each frame. Nevertheless, the latter one is unlikely to be the bottleneck for most real-time applications.

In the scenario of pose recovery, unsupervised dimension reduction techniques like PCA fail to utilize the information of the pose associated with each visual feature vector. Thereby we resort to a dimension reduction algorithm based on supervised spectral embedding to bridge the semantic gap between visual feature space and pose space.

On the other hand, different types of features contribute differently to the overall image similarity. Simple concatenation

* Corresponding author.
E-mail addresses: yujun@hdu.edu.cn (J. Yu), guoyukun@live.com (Y. Guo), dapeng.tao@gmail.com (D. Tao), wanjian@hdu.edu.cn (J. Wan).

http://dx.doi.org/10.1016/j.neucom.2015.04.005
0925-2312/

is not physically meaningful and may be incapable to fully exploit the discriminative power of multi-view features. An alternative solution is to put different weights on different views. We can further consider such weights to be varying depending on the location in the multi-view feature spaces. It is reasonable to assume that the weight is roughly a function of concatenated long vector, i.e., $\mathbf{w}=\mathcal{W}(\mathbf{x})$, which can be approximated in the training stage. The ensemble of NN regressions in the multi-view feature spaces can be done by first determining the weights and then finding the nearest neighbor under the weighted distance measure.

The contribution of this paper is two-fold: (1) a supervised dimension reduction algorithm is proposed to incorporate the information in the pose space for finding the optimal projection; (2) an ensemble approach of nearest neighbor regressions in multi-view feature space is taken to further improve the accuracy of recovery.

2. Related work

2.1. Vision-based pose recovery

After decades of research, there have been a plethora of publications on human pose estimation. Detailed comparison of them is beyond the scope of this paper, so we refer the readers to [5] for a survey of this subject. Using silhouette-based visual features is a very popular choice among published work, including some most significant ones in this field [1,6], as silhouettes can preserve sufficient information for pose estimation while being invariant to clothing and illumination. Silhouette-based algorithms assume that foreground can be reliably extracted, which is true for most indoor scenarios.

The newly developed part-based algorithms, e.g. pictorial structures [7] and poselets [8], have shown very promising results. Part-based methods are capable of recovering highly versatile human poses "in the wild" by making 2D joints/limbs annotations on the pictures. It may seem that those methods have rendered silhouette-based ones outdated, however they still have some disadvantages: (1) they are computationally intensive and commonly cannot operate in real-time; (2) it is nontrivial to lift 2D poses estimated by them to 3D configurations, while silhouette-based approaches can recover the 3D pose directly in a holistic way.

2.2. Dimension reduction

Manifold learning algorithms Laplacian Eigenmaps [9], Locally Linear Embedding [10] and IsoMap [11], as well as their linearized counterparts Locality Preserving Projection [12], Neighborhood Preserving Embedding [13] and Isometric Projection [14] etc., have drawn great attention of researchers recently. Some applications of manifold learning can be found in [15–21]. Specifically, subspace learning techniques have been incorporated to solve pose estimation problems. Closely related to our work, Elgammal et al. [6] explicitly learn activity manifolds along with the mapping functions between visual input space and the 3D body pose space. Chen et al. [4] use trace-ratio criterion to select effective visual feature components. BenAbdelkader applies supervised manifold learning to estimate head pose [22].

2.3. Combining multi-view features

Multi-view spectral embedding algorithms further extend single view methods to make embeddings in different views agree with each other [23,24]. The effectiveness of multi-view learning has been confirmed by its applications in lots of real world problems [25–31,20].

Another way to combine multi-view features is aggregating individual nearest neighbor regressor over all views. Previous work shows that distance metric is crucial for various tasks in machine learning [32–36]. This problem is harder than it appears to be because general ensemble methods like Bagging and Boosting are not directly applicable to k-NN, as they rely on the instability. Unfortunately, most done researches [37–39] in this line are designed with the classification tasks in mind and are not suitable for the regression scenarios. Hence we propose an ad hoc method that better fits our specific problem domain.

3. Proposed algorithms

3.1. Preliminary

The exemplar database consists of the multi-view features of images $\{X^{(v)}\}_{v=1}^{V}$ and the associated ground truth 3D poses $Y=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n]$, where $X^{(v)}=[\mathbf{x}_1^{(v)},\mathbf{x}_2^{(v)},\ldots,\mathbf{x}_n^{(v)}]\in\mathbb{R}^{d_v\times n}$ is the image features from the vth view, and $\mathbf{y}_k=[\cdots,\mathbf{y}_{k^\top j},\cdots]^\top\in\mathbb{R}^{3\cdot M}$ is the XYZ coordinates of all M joints concatenated together.

Denoting the concatenated features as

$$X=\begin{bmatrix}X^{(1)}\\X^{(2)}\\\vdots\\X^{(v)}\end{bmatrix}\in\mathbb{R}^{d\times n},\quad d=\sum_{v=1}^{V}d_v,$$

the goal of supervised spectral embedding is to find a low-dimensional representation $\tilde{X}=[\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2,\ldots,\tilde{\mathbf{x}}_n]\in\mathbb{R}^{d'\times n}$, $d'<d$ with the hope that the Euclidean distances of low-dimensional features can faithfully reflect the (dis)similarities of the poses. Afterwards, the inference of the novel pose $p(\mathbf{y}|\mathbf{x})$ can be deduced to k-nearest neighbor regression with low-level feature $\tilde{\mathbf{x}}$ in the database of training exemplars. The overall framework is illustrated in Fig. 1.

3.2. Visual feature extraction

Visual feature is extracted from the silhouette for robustness. Following [4], we choose 5 feature descriptors in total,[1] which have proved to be efficient in the computer vision literature:

(1) *Occupancy map*: Crop the silhouette according to its bounding box, and divide it by 12×8 grids. Calculate the percentage of foreground pixels inside each grid, and concatenate those values together columnwise. This generates a 96D vector, each element of which ranges from 0 to 1.
(2) *Contour signature*: Start from the topmost point, uniformly sample 64 points along the contour of the silhouette. Under three variations of measures, coordinates, distances to the centroid, and tangent angles, we get a $128+64+128=320$ dimensional feature.
(3) *Fourier descriptor*: We treat the output feature of contour signature above as discrete signal, and apply DFT (Discrete Fourier Transform) to it. Eliminating the DC component results in a feature of $62+32+62=156$D.
(4) *Hu moments*: Compute 7 image moments in Hu's invariant set, by considering the shape as either contour or area, a $7\times 2=14$ dimensional feature is obtained.
(5) *Shape contexts*: We use shape contexts containing 12 angular bins and 5 radial bins. Construct a codebook with the size of 100 by running k-means clustering on the generated 60D features, and quantize them to form the 100D histogram of shape contexts.

The feature vectors are normalized to prevent the dimensions with large ranges overpowering others.

[1] We do not use *Poisson features* because it is very time-consuming to compute.