International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.9 (2014), pp.1-8
http://dx.doi.org/10.14257/ijmue.2014.9.9.01

Motion Recognition based on Sparse Representation and 3D Spatial–Temporal Feature

Jian Xiang

School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China
freexiang@gmail.com

Abstract

The emergence of a large number of databases for capturing 3D human motions has made the efficient analysis and processing of human motion data to effectively use these databases a new challenge. To reduce high-dimension complexity, a dimensional feature based on the 3D spatial–temporal characteristic should be extracted from human motions. Moreover, the motion data should be re-expressed by sparse representation to realize the projection from high dimensional data to a low-dimensional subspace. The different motions should then be recognized and classified to obtain the automatic recognition and automatic retrieval of 3D human motions.

Keywords: *spatial–temporal feature; sparse representation; subspace; classify*

1. Introduction

Currently, the analysis on 3D human motion data lacks a complete and effective analysis and processing technique. Therefore, the large-scale 3D human motion databases cannot be applied to the field of digital media efficiently, quickly, automatically, and intelligently.

3D human motion data contain many semantics, such as object, event, behavior, and scenario. Their characteristics of magnanimity, non-structure, high-dimension, and multistage greatly challenge semantic comprehension.

In recent years, the combination of compressed sensing theory and variable selection method (henceforth feature selection in this application when high-dimensional data (*e.g.*, image) are analyzed) has been applied to form a more effective "sparse representation" of media data, which have become a research hotspot in the fields of computer vision and machine learning, among others. Compressed sensing adopts the priori knowledge that "data are sparse and can be compressed" to reconstruct signal. David Donoho and Emmanuel Candes from Stanford University and Terence Tao from University of California, Los Angeles, conducted several representative studies involving stochastic matrix, signal restoration, and sparse measurement, among others [1-2].

The analysis of 3D human motion data to recognize actions is a challenging problem. Matching multi-dimensional spatio-temporal movement patterns against large-scale 3D human motion databases is difficult to achieve efficiently. 3D human motion data involves context such as appearance, background, viewpoint, event, behavior and scenario, and motion recognition typically requires interpretation of unstructured, high-dimension and multistage feature sequences. Until now, several motion features have been proposed: [3] introduce the Energy-based Least Square Twin Support Vector Machine (ELS-TSVM) algorithm for human action recognition which can handle unbalanced datasets' problem. The combined saliency of motion and appearance based on kernel histogram [4] is used in human action recognition. [5] have extended the method of selecting the most discriminative features using

ISSN: 1975-0080 IJMUE

the AdaBoost algorithm to the human-action recognition task. [6] builds Gaussian mixture model(GMMs) based on the views of each object and then accomplishes 3D retrieval and recognition based on the distance between GMMs.

In the course of data analysis, when the sample feature dimensions are far greater than the sample numbers, the conventional approach is unable to forecast and identify the data precisely. Tibshirani of Stanford University and Breiman of University of California, Berkeley, put forward least absolution shrinkage theory and selection operator theory, respectively, nearly at the same time to pose 11-norm constraints on the characteristic coefficient, prompting the greatest sparseness of the selected feature to guarantee result stability and improve interpretability [7-8]. The variable selection method represented by Lasso has become the main means to analyze the high-dimensional data of statistics. Currently, related research focuses on how to design a better target optimization function, more interpretable regularization factors, and more effective solution algorithms [9].

In consideration of the advantages of compressed sensing and variable selection in data processing, Wright and Ma Yi from the University of Illinois, Urbana Champaign, introduced them into face recognition and proposed the new thinking of conducting face recognition through l1-norm constraints [10]. As media data have many extractable features, finding a way to search for an effective sparse representation from high-dimensional features and then to study the theory and method of semantic understanding of media data has become a development tendency of the current computer vision and pattern recognition field applied to visional word selection [11], image annotation [12], and image restoration [13], among others. In terms of recognizing the real world, researchers from the NEC California Laboratory and the research group of Prof. Thomas S. Huang from the University of Illinois, Urbana Champaign, collaborated to apply sparse representation to visual object recognition, winning them the first prize in the PASCAL Visual Object Classification Challenge [14].

In this paper, we introduce the basic methods and approaches of motion semantics based on the sparse representation mechanism of the 3D spatial–temporal features (SRSF) and focus on solving the rapid analysis and interpretability problems of 3D human motion data.

2. Extraction of Motion Data Features

Regard the captured human motion data M as the human posture sequence obtained through sampling at discrete time point. Every sampling point is a frame and the posture of every frame is determined by 16 articulation points jointly. In this way, at random frame time i, the human posture is expressed as: $F_i = (p_i^{(1)}, r_i^{(1)}, r_i^{(2)}, ..., r_i^{(16)})$, in this equation: $p_i^{(1)}$ Î P^3 and $r_i^{(1)}$ Î R^3 represent the place and direction of Root articulation point respectively, i.e. translation quantity and rotation quantity; $r_i^{(j)}$? $R^3, j \quad 2...16$ represent the direction (rotation quantity) of non-Root articulation point. According to the interrelation of all articulation points of human skeleton, at random frame time i, the place of random non-Root articulation point N_j of human skeleton can be obtained through formula (1) through 3D transformation formula:

$$\vec{p}_i^{(j)} = T_i^{(root)} R_i^{(root)} ... T_0^{(grandparent)} R_i^{(grandparent)} (t) T_0^{(parent)} R_i^{(parent)} \vec{p}_0^{(j)} \qquad (1)$$

In this formula, $\vec{p}_i^{(j)}$ represents the world coordinate of articulation N_j at time i; $T_i^{(root)}, R_i^{(root)}$ represent the translation and rotation transformational matrix of Root articulation point at time respectively which are created by $p_i^{(1)}, r_i^{(1)}$; $T_0^{(k)}$ represents the translation transformational matrix of N_k (N_k is the random node from the root node and N_j node) generated by the translation quantity of the local coordinate system of its parent articulation point at initial time; $R_i^{(k)}$ represents the rotation transformational matrix of articulation point N_k (value of N_k is the same as above) which is generated by r_i^k ; $\vec{p}_0^{(j)}$ represents at initial time, the translation quantity of N_j under the local coordinate system of its parent articulation point.

This article extracts the frames with same local spatial feature at single articulation point to combine and form a space set on the basis of extraction of spatial-temporal feature to extract a keyspace which can represent transformation law of motion space, thus realizing reduction of the first step for original motion data. While based on the independence of spacial feature of every articulation point, the index lists will be established for 16 articulation points of human body which are for calculating the local similarity of every articulation point. Then the calculated amount will be reduced greatly comparing with DTW algorithm. In order to obtain contributions made by the all index lists nodes during comparison of motion similarity, the way of a serial of machine learnings can by adopted, for example, a decision tree can be get through learning and training of the original motion data of data driven decision tree; the nodes of decision tree are properties of the articulation points. The layer is higher, the impact of this node on the motion matching result is greater. In this way, we can learn the impact degree of every articulation point on overall human motion, thus parameterizing the weight of articulation point's impact on the motion. Then conduct similarity calculation for the motion example *Q* and index list of the articulation point with largest weight of motion *A* in motion base. If the result is not similar, *A* could be passed and conduct similarity calculation for motion *B* directly. Only when the similarity of the articulation point with larger weight is within range of a valve-value can we compare the subsequent articulation point. Large parts of meaningless calculations can be avoided in the process of retrieval.

2.1. Calculation of Three-dimensional Space

We calculate the world coordinate of every articulation point and get a 51-dimensional data through formula 1. Removing the *Root* articulation point, 16 articulation points, 48-dimensional data will be obtained.

The motion is expressed as:

$$\begin{gathered} M_s = (F_1, F_2, \ldots, F_i, \ldots, F_n) \\ F_i = (p_{i1}, p_{i2}, \ldots, p_{ij}, \ldots, p_{i16}) \\ p_{ij} = (x, y, z) \end{gathered} \tag{2}$$

In these formulas, *n* is frame number of motion data; p_{ij} is the world coordinate of the *j* articulation point of the *i* frame.

We utilize this 48-dimensional data to generate spatial alteration of every articulation point. Firstly, define the spacial sets S_{up} and S_{down} respectively for upper half and bottom half of human body, S_{down}, S_{ki} ? $S_{up}, i \quad 1, 2, \ldots m$; S_{lj} ? $S_{down}, j \quad 1, 2, \ldots, m$; *m* are spacial number of the spacial sets. Now we will divide the upper half and bottom half into equal spacial sets and S_{ki}, S_{lj} is independent space of upper and bottom spacial sets. Regard *Root* as benchmark and make the articulation points above *Root* nodes correspond to S_{up} and the articulation points below *Root* nodes correspond to S_{down}. When the articulation points of upper limb, the spacial transformation will correspond to the value of S_{ki}.

Several partition rules will be defined below:

$$front\ (N_i, N_j) = \begin{cases} 1, N_i \text{ in front of } N_j \\ 0, N_i \text{ behind of } N_j \end{cases} \qquad left\ (N_i, N_j) = \begin{cases} 1, N_i \text{ left to } N_j \\ 0, N_i \text{ right to } N_j \end{cases}$$

$$high\ (N_i, N_j) = \begin{cases} 1, N_i \text{ above } N_j \\ 0, N_i \text{ below } N_j \end{cases} \qquad far\ (N_i, N_j) = \begin{cases} 1, N_i \text{ distance from } N_j > \lambda \\ 0, N_i \text{ distance from } N_j < \lambda \end{cases}$$

Define the spacial transformation of motion $B = (b_1, b_2, \ldots, b_n)¢, b_i = (s_{i1}, s_{i2}, \ldots, s_{i16})$. b_i is the spacial transformation of articulation point *i*; s_{ij} represents spacial transformation of frame *j* of articulation point *i*. Assume s_{aj} represents spacial transformation of articulation point *a* of the upper limb:

The Table 1 is the spacial rules of spacial transformation of s_{aj} created using the above definition rules.

Table 1. Space Rules

s_{aj}	*Front* (N_{a1}, N_{ar})	*Left* (N_{a1}, N_{ar})	*High* (N_{a1}, N_{ar})	*Far* (N_{a1}, N_{ar})
$s_{aj} = s_{k1}$	1	1	1	1
$s_{aj} = s_{k2}$	0	1	1	1
…	…	…	…	…
$s_{aj} = s_{km}$	0	0	0	0

While rule of $front, left, high$ is obtained through 48-dimensional data of formula (1). As the calculation rules are among different nodes of the same frame, the calculation amount is small. Therefore we calculate all articulation points and obtain the spacial transformation of human body from the initial motion data. Every locality of this spacial transformation directs at every articulation point and is relatively independent.

2.2. Sparse Representation of Motion

During analysis and processing of motion data, we use a n-dimensional vector $m \in R^n$ to represent a motion sequence. The vector here can be obtained through arranging all articulation point data in sequence and can be a feature vector of motion. In this way, assume that we have the y sequences of different roles of a motion, such as $m_1, m_2, \ldots m_y \in R^n$. For a new motion sequence $m \in R^n$.

$$m = \sum_{i=1}^{y} \beta_i m_i \tag{3}$$

In this formula, β_i is linear representation coefficient to be written as matrix form:

$$m_{y \times 1} = T_{y \times m} x_{m \times 1} \tag{4}$$

X is coefficient vector. Then we will present the sparse representation of motion based on this compactness representation. Under the practical circumstance, what stored in the database is the multi-roles of all motions. In this section, we will give the global representation of the motion sequence to be recognized in the overall database.

Assume there are k motions in database, the No. i motion has y_i motions of different roles. For the No. i motion, it will extract yi n-dimensional feature vectors which are $m_1, m_2, \ldots m_y \in R^n$. The first subscript i represents the No. i motion and the second subscript j is its No. j motion sequence. Then we will adopt matrix to represent these vectors: $T_i = [m_1, m_2, \ldots m_y] \in R^{n \times y_i}$. Thus, correspond a matrix T_i to every motion sequence. If there are k motion sequences, there will be k matrix: $T_1; T_2; \ldots; T_k$. Connect these matrix series to get a large matrix of all motions in the overall database:

$$T = [T1; T2; \ldots Tk] \in R^{n \times y} \tag{5}$$

Now we are considering the global representation of a motion sequence to be recognized. Assume the to-be-recognized motion comes from No. i motion sequence and its feature is f, then put this motion sequence in the global database through equation (6):

$$f = \sum_{i=1}^{y_j} \beta_i m_{j,i} = \beta_1 m_{j,1} + \ldots + \beta_{y_j} m_{j,y_j} \tag{6}$$

The contributions made by the same motion will be listed as the above equation and different people makes 0 contribution, then we will have:

$$F = 0 \cdot m_{1,1} + \ldots + 0 \cdot m_{j-1,y_i-1} + \beta_1 m_{j,1} + \beta_2 m_{j,2} + \ldots + \beta_{y_j} m_{j,y_j} + 0 \cdot m_{j+1,1} + \ldots + 0 \cdot m_{k,y_k} \tag{7}$$

The above formula is the global representation of the motions to be recognized in the data base. It is easy to write as matrix form:

$$f_{n \times 1} = T_{n \times y} x_{y \times 1} \tag{8}$$

We point out that when there are numerous motion types in the database, namely k is relatively large, the linear representation given by equation (8) is sparse. As there are only yi nonzero elements in the y-dimensional vector x and $y_i / y \approx 1/k \ll 1$, namely $n_i \ll n$. In other words, the nonzero elements occupy small parts in vector x. So far, we have shown the sparse representation of the motion to be recognized.

For the motion input, when it is represented by global motion linear, only one motion will make greater contributions to different sequences of juese and contributions made by the rest motions are nearly zero (o is inexact as there will be errors).

2.3. Motion Recognition based on Sparse Representation

In the course of practical motion recognition operation, a database containing multi-roles motion sequence is given and input a motion sequence to be recognized. It requires deciding what kind of motion in the database this sequence belongs to. But in practical situation, as the motions are multiple in the database, namely n is very large. While considering the complexity of calculation, the valuation of dimension degree of motion features will not be very high, namely m is relatively small, then under normal circumstances, m < n or even m ≪ n. Now the equation is a indefinite equation, *i.e.* the unknown number (far) more than number of equations and the equation is unable to obtain a unique solution. In this case, it require adding certain limiters to make solution problems of equation become a optimization problem. The common limiter is minimization of ℓ_2 form, *i.e.* least square method.

However both theory and experiment prove that least square method is not favorable for the sparse representation. In fact, the reference [15] has further proven that the limiter given by `0 norm minimization is optimal, *i.e.* Do not transform the solution of equation (8) to the next optimization problem:

$$\arg\min_{\beta} \|\beta\|_0 \quad \text{subject to } X^{'}\beta = y \tag{9}$$

As it is an NP-hard problem which cannot be solved directly. Based on the compression sensing theory raised by Donoho, if the signal is sparse enough, then the solution of ℓ_0 minimization problem is equal to solution of ℓ_1 minimization problem. The unique sparsest solution can be obtained through minimizing ℓ_1 norm and transform to the following lasso problem:

$$\arg\min_{\beta} \frac{1}{2}\left\| y - X^{'}\beta \right\|_2^2 + \lambda \sum_{j=1}^{K} \sum_{i=1}^{n_i} \left| b_{ji} \right| \tag{10}$$

The compressional sampling theory can prove that optimization problem given by equation (10) can approach equation (9) in a limit with probability of 1, i.e. the error probability of these two optimizations approximate 0. Therefore, what we solve in fact is the ℓ_1 norm optimization problem of equation (10). So the solution of optimization problem of ℓ_1 minimization can be obtained through solving Lasso problem.

3. Experimental Results and Analysis

To compare the recognition efficiency of the classifier based on spatial-temporal features and sparse representation (SRSF) over common motions, we adopt common 25 motion types to make analysis. While as comparison, the most representative retrieval method based on clustering hierarchical retrieval tree (CIT) will be applied as reference. The results are shown in Figure 1 below. Apparently the retrieval of common motion is quicker and preciser. Table 2 shows the time comparison between 2 retrieval methods. Table 3 indicates the retrieval precision of 2 methods. In addition, the speed and efficiency of CIT depend on the scale of the database. With expansion of database scale, the time for retrieval will increase greatly, but the recognition efficiency of this method is irrelated to scale of database basically. Because the sparse representation has greatly reduce the dimension degree of motion data and calculation complexity, the recognition speed will then be faster.

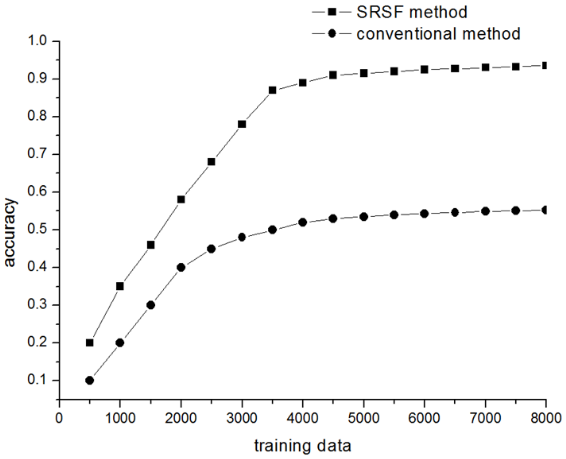


Figure 1. Comparison of the Performance of Conventional Method with SRSF

Table 2. Recognition Time

Motion clips	Recognition time(second)			
	Motion recognition by CIT		Motion recognition by SRSF	
	N=200	N=800	N=200	N=800
A(52)	1.6325s	4.5467s	0.7189s	0.8134s
B(101)	2.0115s	5.5578s	0.8235s	0.9102s
C(163)	1.9456s	5.3413s	1.2381s	1.3455s
D(260)	2.0945s	6.5984s	1.3453s	1.5035s

Table 3. Recall and Precision

Motion clips	Recall		Precision	
	CIT	SRSF	CIT	SRSF
walk	0.79	0.96	0.88	0.95
run	0.75	0.95	0.85	0.97
jump	0.52	0.91	0.71	0.90
bunch	0.49	0.89	0.45	0.88

Table 4 analyzes the direct original data processing and compares the two methods of conducting retrieval and processing after data dimensional degree reduction through traditional ISOMAP dimensionality reduction and sparse representation. The results show that our recognition efficiency is far higher that direct processing and the time for training is less than the traditional dimensionality reduction.

Table 4. Training Time

Motion data	Training time(second)			
	Walk	run	jump	bunch
Original motion feature	65.1145s	63.2135s	80.9145s	97.1394s
ISOMAP D-data	9.4985s	10.2571s	12.6590s	14.6914s
Data by SRSF	5.9981s	7.4510s	9.1240s	10.1198s

Acknowledgements

This work is based upon work funded by Natural Science Foundation of P.R. China under Grant No 51246002 and Zhejiang Provincial Natural Science Foundation of China under Grant No. Y1080366, Y1100219

References

[1] D. Donoho, “Compressed Sensing”, IEEE Transactions on Information Theory, vol. 4, no. 52, **(2006).**

[2] E. Cades and T. Tao, “Reflections on compressed sensing”, IEEE Information Theory Society Newsletter, vol. 4, no. 58, **(2008).**

[3] J. A. Nasiri, N. M. Charkari and K. Mozafari, “Energy-based model of least squares twin Support Vector Machines for human action recognition”, Signal Processing, vol. 104, **(2014).**

[4] Y. Yang and Y. K. Lin, “Human action recognition with salient trajectories”, Signal Processing, vol. 93, **(2013).**

[5] L. Liu, L. Shao and P. Rockett, “Human action recognition based on boosted feature selection and naive Bayes nearest-neighbor classification”, vol. 93, **(2013).**

[6] M. Wang, Y. Gao, K. Lu and Yong Rui, “View-Based Discriminative Probabilistic modeling for 3D Object Retrieval and Recognition”, IEEE Transactions on Image Processing, vol. 4, no. 22, **(2013).**

[7] R. Tibshirani, “Regression shrinkage and selection via the lasso”, Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 1, no. 58, **(1996).**

[8] L. Breiman, “Heuristics of instability and stabilization in model selection”, The Annals of Statistics, vol. 6, no. 24, (**1996**).

[9] J. Fan and J. Lv, “A selective overview of variable selection in high dimensional feature space”, Statistica Sinica, vol. 20, (**2010**).

[10] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma, “Robust face recognition via sparse representation”, IEEE Transactions on Pattern Analysis and Machine intelligence, vol. 2, no. 31, (**2009**).

[11] J. Yang, K. Yu, Y. Gong and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification”, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, (**2009).**

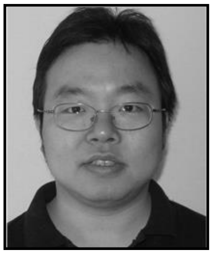
[12] D. Hsu, S. Kakade, J. Langford and T. Zhang, “Multi-label prediction via compressed sensing”, Proceedings of Advances in Neural Information Processing Systems, (**2009).**

[13] J. Mairal, M. Elad and G. Sapiro, “Sparse representation for color image restoration”, IEEE Transactions on Image Processing, vol. 1, no. 17, **(2008).**

[14] Y. Gong, T. Huang, F. Lv, J. Wang, C. Wu, W. Wu, J. Yang, K. Yu, T. Zhang and X. Zhou, “Image classification using Gaussian mixture and local coordinate coding”, The PASCAL Visual Object Classes Challenge Workshop, (**2009).**

[15] D. Donoho, “For most large underdetermined systems of linear equations the minimall 1-norm solution iS also the sparsest solution”, Communications on Pure and Applied Math, vol. 6, no. 59, **(2006).**

Author



Jian Xiang, he received his Ph.D. degree from the College of Computer Science and Technology, Zhejiang University. He has been associate professor at Zhejiang University of Science and Technology. He is a member of the China Computer Federation. His research interests include multimedia analysis and retrieval, computer animation and statistical learning, etc.

E-mail:freexiang@gmail.com