International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.7 (2015), pp.299-308
http://dx.doi.org/10.14257/ijsip.2015.8.7.29

Generative Human Action Tracking Based on Compressive Sensing

Gaofeng Li[1], Fei Wang[1] and Wang Lei[1,2]

[1]*School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*
[2]*Chinesisch-Deutsches Hochschulkolleg, Tongji University, Shanghai 201804, China*
2011gaofengli@tongji.edu.cn, 1110373@tongji.edu.cn, leiwang@tongji.edu.cn

Abstract

Action tracking and recognition is a challenge due to human deformation and complex scene system. Tracking-by-detection methods are used to solve appearance changes problem caused by viewpoint, occlusion, scale or deformation. Here we propose a robust object tracking and generative action recognition method. Compressive sensing is improved to track object with superpixels, and the generative structural part model is designed to be adaptive to variation of deformable object. We evaluate the method on challenging sequences. Also, we make qualitative and quantitative discussion. The results indicate the method is robust, and it is adaptive to deformable object tracking and action recognition.

Keywords: *object tracking, action recognition, compressive sensing, generative part model*

1. Introduction

Human action recognition is the process of labeling image sequences with action words, which has main applications for video surveillance, human machine interaction and autonomous navigation in computer vision [1, 2]. Object tracking and action recognition are respectively spatial and temporal process. Robust tracking methods are proposed to learn and update model of object with appearance changing. However, it is difficult to track the deformable object, articulated parts of which produce complex relative motion. We merge a generative tracking conception into action recognition and combine temporal feature invariance with spatial structural integrity.

Object tracking generates an inference about the motion of object given a sequence of video and images [3]. The tracking methods are divided into two categories: generative method and discriminative method. The generative method makes a dynamic model to estimate the motion of object, which is bottom-up and proper to rigid object. The discriminative method poses candidates matching as binary classification problem, known as tracking-by-detection. The top-down online trackers tend to drift when the appearance changes much. Human tracking might be treated as action tracking whenever human is always acting or in static posture. The deformation for human action changes overall appearance largely, but is utilized under generative structural model we proposed.

Human action analysis methods could be categorized into three major classes: non-parametric, volumetric, and parametric time-series approaches [4]. The non-parametric method extracts object features in each frame and conducts template matching. The templates of posture demand massive offline training and computation for recognition. The method is typically used to static posture estimation, such as DPMs by Felzenszwalb [5]. The volumetric method considers an action as a 3-D volume of pixel intensities with continuous frames. This method is actually based on spatial and temporal filters, which require effectively capturing the action models. The parametric time-series method models

ISSN: 2005-4254 IJSIP

the temporal dynamics of the motion, which parameters for a class of actions is calculated with known action training data. The action is presented as continuous state sequences with the probability statistical model, such as Hidden Markov Model (HMM) [6], Linear Dynamical Systems (LDS) [7] and Dynamic Bayesian Network (DBN) [8]. The action referred in the methods is regular movement under constrained settings. The real action is changeable at each viewpoint and different from standard action in discriminative training model.

In this paper, we propose a generative action tracking algorithm, which combines generative structural model and object tracking with compressive sensing. The object is featured by local salient parts in compressed domain and represented by structural integrality. Generative structural model is developed to track human and analyze action in real world.

The rest of the paper is organized as follow: In Section 2, compressive sensing is reviewed and improved by superpixels. Section 3 gives a detailed description of generative structural object model. In Section 4, analysis and discussion are performed with our method and state-of-the-art trackers. The conclusions are drawn with advantages and suggestions for future research in Section 5.

2. Compressive Representation

Object tracking is generally composed of representation, model and updating. The representation could be point, shape or silhouette of object, from which the features are extracted with intensity, color and texture. Compressive sensing is an efficient data processing method, which is employed to extract features of object with sparse measurement matrix [9]. Compressive sensing compresses massive data to decrease computation in image and video processing.

For meeting the conditions of sparsity, incoherence and RIP (restricted isometry property), a high-dimensional x could be projected to low dimensional space with random measurement matrix [10]:

$$x = \Psi\alpha, \tag{1}$$

$x \in R^N$, Ψ is the base vector space, α is the sparse representation of x in Ψ domain. The measuring function is given as

$$y = \phi \cdot x = \phi \cdot \Psi\alpha = \Phi\alpha, \tag{2}$$

$\Phi \in R^{M \times N}$ is random measurement matrix, $M << N$.

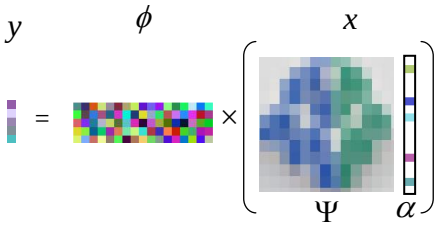
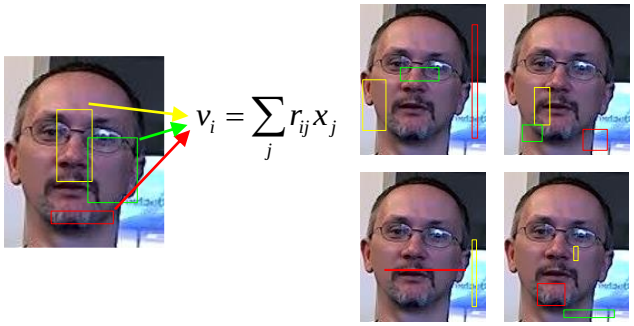


Figure 1. Compressive Sensing Theorem

We deploy the structural parts of object to constrain sparse representation in compressive sensing (CS). The sparse representation is combined with object features for great capacity of object recognition. Figure 1 demonstrates the fundamental theorem of CS in 2D. For image data, random Gaussian matrix satisfies Johnson-Lindenstrauss lemma and RIP [10]. So the measurement matrix is produced by random Haar-like rectangles [9], as shown in Figure 2. For the sake of tracking object with appearance changing, the position and size of

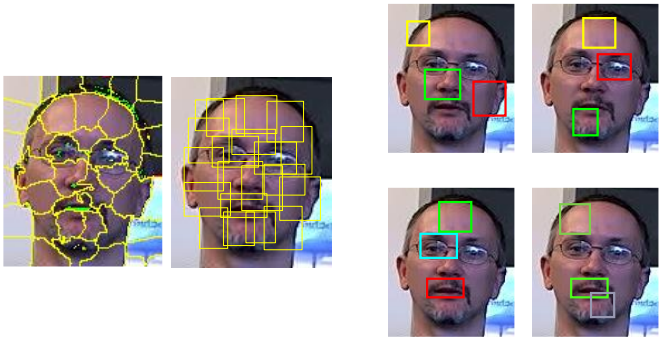
rectangles should be constrained according to the features of object. We reconfigure the random rectangles with structural conditions by superpixel method SLIC (Simple Linear Iterative Clustering) [11]. The object is segmented into superpixel clusters, which are randomly grouped by three or four rectangles to form measurement matrix as shown in Figure 3. The modified rectangles are close to local salient features of object.



(a) Compressive representation **(b) Random Haar-like rectangles**

Figure 2. Compressive Sensing

The comparison between the two compressive representations is shown with normalized response curve of object tracking in Figure 4. The axis x is sample data; the axis y is normalized response value, in which the maximum is accordingly possibly position of object. The red dot line denotes the response value of samples for the constrained rectangles (see top right sample in Figure 4). And the blue cross line is the value for general rectangles (see top left sample in Figure 4). The improved sparse measurement matrix responses sharply and has strong discrimination from samples.



(a) SLIC clusters **(b) Grouped rectangles**

Figure 3. Compressive Sensing with Superpixels

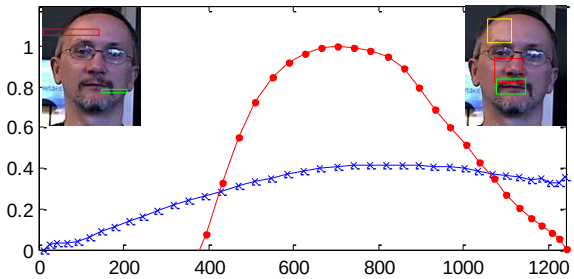


Figure 4. Response Curves of Compressive Representations

3. Generative Structural Method

Part model is stemmed from pictorial structures by Fischler [12], made of local feature parts and springs connecting parts as their location relationship. Felzenszwalb and coworkers [5] proposed deformable part model for global minimizing function and fast detection in order to avoid unnecessary decisions. Part-based model detects object and parts in one image with a large number of samples training offline. For human action, we propose the generative structural tracking model, which combines with improved compressive tracking and generative structural relationship of parts for action tracking.

3.1 Structural Model

The structural model consists of root and parts. The root estimates the overall object with constrained compressive sensing. The salient parts meet the structural relationship and constraints for object. The objective function is minimized to locate the root and parts including root filter $H(v)$ and parts energy function $E(p)$.

$$score = H(v) + E(p) \tag{3}$$

The root filter is deployed with Bayesian formula for samples:

$$H(v) = \log\left(\frac{p(y_1 \mid x)}{p(y_0 \mid x)}\right) = \log\left(\frac{\Pi_{i=1}^{n} p(y_1 \mid v_i)}{\Pi_{i=1}^{n} p(y_0 \mid v_i)}\right), \tag{4}$$

v_i is the set of representation for sample x in compressive domain, y_1, y_0 are positive and negative samples.

The energy function of parts consists of an appearance term and a smoothing term. The appearance term computes matching degree between part features and its template. The smoothing one penalizes the spring deformation of each part including distance and orientation.

$$E(p) = \sum_{i=1}^{n} F_i \phi(x, y) + \sum_{i,j}^{n} P_{ij} d(x, y), \tag{5}$$

F_i is each part filter with template, which computes each part matching degree. $\phi(x, y)$ is the feature vector of parts. P_{ij} is the spring relationship between each other, which calculates deformable degree. $d(x,y)$ is a set of distance of parts. The relationship of parts restricts the changes of parts from distance to orientation. When the matching degree of one part is high, their relationship is weighted.

The root and part filters are merged into each other. The root filter obtains a coarse object. The part filters refine the root result and avoid wrong detection because of appearance changing, such as occlusion, fast motion, rotation and deformation. The parts result is fed back to root filter for achieving an optimal tracking.

3.2 Generative Parts Method

The optimization of objective function is time-consuming on sequence images. We develop the simple human parts for tracking. The body model is consisted by head, torso, hip, limbs, as shown in Figure 5. The generative parts are grouped into two types. Head and torso are main parts of human and the features of them are stable in parts I. They are independently tracked. The limbs are main deformable components and attached to head and torso as parts II, which are passively matched with human parts model. All the parts features and spatial structure are extracted to calculate objective function.

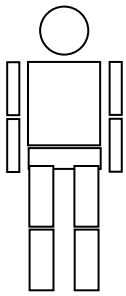


Figure 5. Human Parts Model

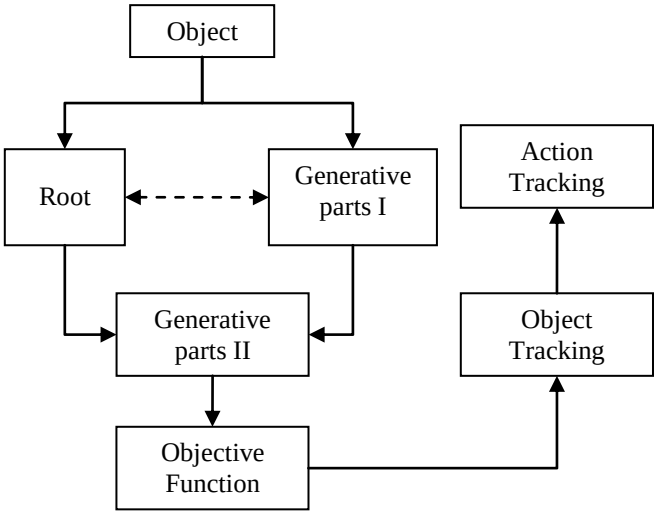


Figure 6. Generative Structural Method

The outline diagram is presented in Figure 6. Root is the overall tracking module, which confidence reflects the deformation degree of the entire appearance. Generative parts I accordingly tracks head and torso parts. When object changes a lot in appearance, root response becomes low. The high confident generative parts are used to relocate root to filtering samples. When the response is high, the generative parts I is fine tuned. From the synthetic result, the generative parts II are recognized with the structural parts model. Object and action tracking is processed after minimizing objective function.

3.3 Algorithm Details

The moving object is first detected by the detector of scene. The human object is initialized with general human body model. When *i*-th frame is input, the root, head and torso are respectively detected with compressive presentation from samples in searching areas. According to the spatial structural model in last frame, the structural object is located with root, head and torso as maximum likelihood estimation. The $d(x,y)$ and P_{ij} are adjusted to compute $E(p)$ in Eq. (5). During limited iterations, the score is minimized to confirm probable object in Eq. (3). The detail of the algorithm is shown as follows:

Initialize: root template and parts model with SLIC.
Input: *i-th* frame I_n, filter F_i, parts filter P_{i-1}, $d(x_{i-1},y_{i-1})$.
(1) Sample sliding windows around the previous location (x_{i-1}, y_{i-1}) in multi-scale and extract features with compressive representation v_i.
(2) Filter samples with root model $H(v)$ to calculate response c.
(3) Filter head/torso samples in searching windows.
(4) If $c > T_{\text{threshold}}$,
  a. Confirm parts location and extract features from corresponding samples.
  b. Update the structural relationship between root and parts.
(5) Otherwise,
  a. Decrease the confidence of root and confirm the head/torso parts matching degrees.
  b. Reconfigure the constraints of generative structural model.
(6) Estimate and detect the limbs parts.
(7) Extract parts features and compute $E(p)$ with parts filter P_{i-1}, $d(x_{i-1},y_{i-1})$.

(8) Repeat to step (2), until minimize Eq. (3).
(9) Update root and parts model with linear learning rate λ .
Output: tracking results in the current frame.

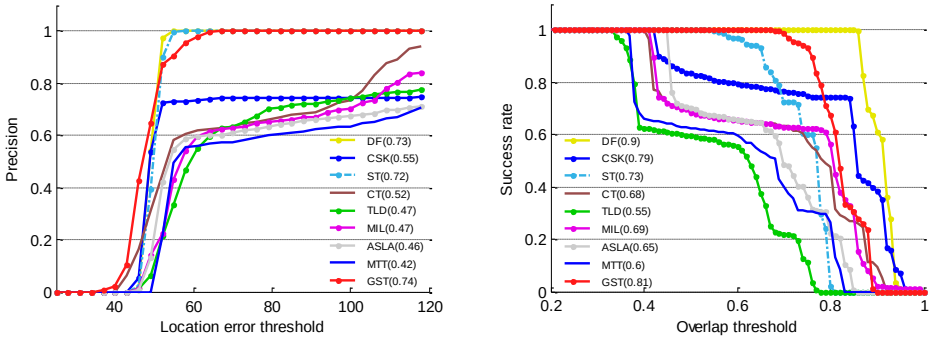
The algorithm is actually tracking-by-detection tracking based on generative method. It provides an efficient and robust tracking method to avoid complex discriminately learning offline for deformable object. The generative parts structure adjusts to real object not confined to training images. The generative structural algorithm optimizes human object tracking and analyzes action.

4. Results and Analysis

We conduct experiments of generative structural tracking model on real world videos. The proposed algorithm performs well for object and action tracking under different deformations. The quantitative and qualitative analyses are carried out for the experimental results.

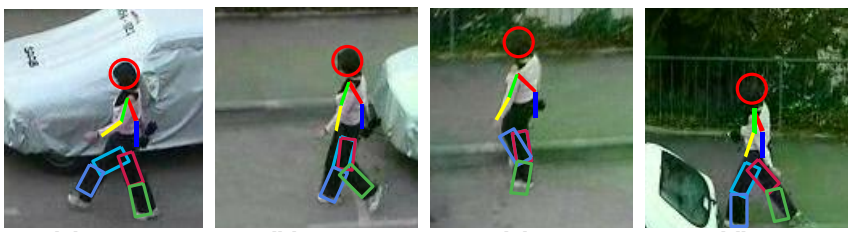
We test our method GST on challenging sequences from standard tracking dataset [13] and compare the precision and success rates with state-of-the-art tracking algorithms. Figure 7 presents the precision and success of trackers on a sequence named Woman from the dataset. The numbers in brackets are the average value of precision and success. The GST is robust and adaptive to overall object tracking comparable to state-of-the-art trackers.

Moreover, GST tracks the human action based on the parts of deformable object. In frames of Figure 8, the walking woman is tracked and recognized with the generative parts structure. We establish spatio-temporal action template with probability density distribution from continuous tracking frames. The frequency of parts occurrence determines the action pattern. Figure 9 shows the probability density distribution of walking on the sequence. High probability values are scattered among the torso and legs fields, which characterize walking. We implement the means to analyze and recognize action.



(a) Precision **(b) Success**

Figure 7. Precision and Success of Sequence Woman



(a) 16 **(b) 43** **(c) 81** **(d) 100**

Figure 8. Human Action Tracking on Sequence Woman

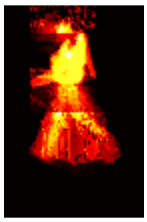


Figure 9. Probability Density Distribution of Walking

We build a high-definition dataset for object tracking and action analysis. The dataset is an indoor scene including single and multiple people. The sequences are annotated with action labels include walking, crouch, sitting, bowing, jogging, talking, and etc. The bounding boxes for tracking are sampled spatially and temporally to evaluate the robustness of trackers.

We qualitatively evaluate GST for this dataset. The method is robust to human object tracking and adaptive to action. As shown in Figure 10, GST is effective and efficient to track human object under different kinds of actions. The crouch action shows that the legs change largely but head and torso a little. Sitting is between standing and crouch in structural changes. The torso is falling down while bowing. The jogging only changes limbs posture. GST still tracks people interaction. The experimental results are satisfied for monocular camera.

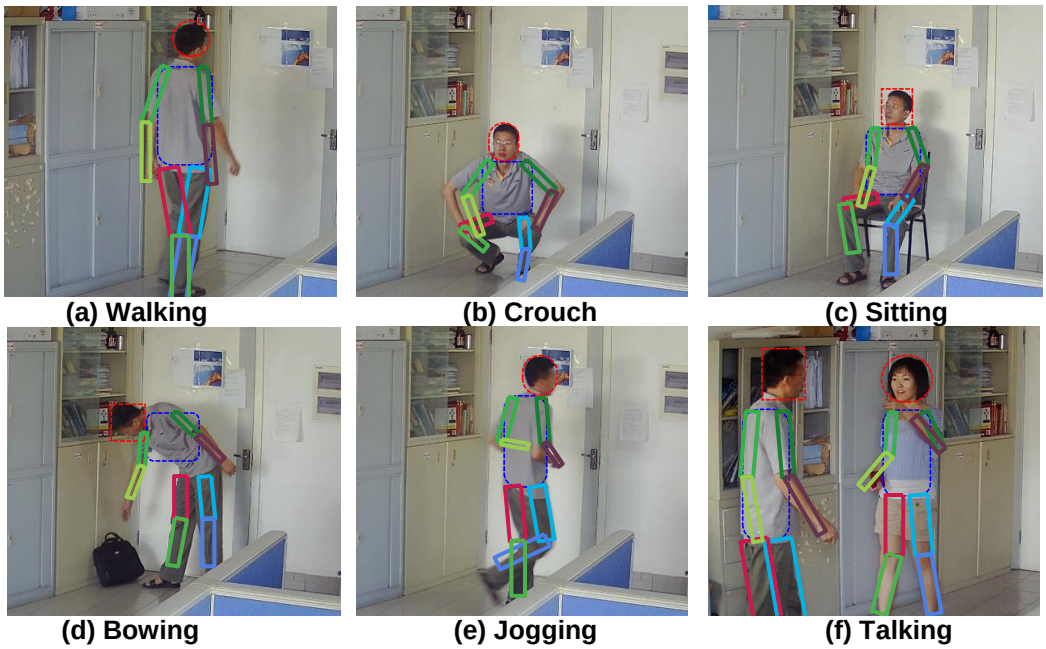


Figure 10. Sequence Results

5. Conclusions

Generative structural tracking is a novel tracking method, which is not only comparable to state-of-the-art trackers, but also robust to human tracking and action analysis. GST improves compressive sensing to increase capacity for object tracking. The method treats human as root and two type parts, which are adaptive to deformable object and complex scene. Parts are the reason of deformation, but they are helpful to track the whole object in generative structural model. The view point is a disturbance factor impacting on GST. In future work, we will further study action tracking and recognition in multiple views.

Acknowledgements

This work is partially supported by National High-tech Research & Development Program (No.2012AA7041003). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

References

[1] R. Poppe, "A survey on vision-based human action recognition", Image and Vision Computing, vol. 28, no. 6, **(2010)**, pp. 976-990.
[2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition", Vision, Perception and Multimedia Understanding, **(2010)**,
[3] D. A. Forsyth and J. Ponce, Computer Vision: A Modern Approach, 2E. New Jersey: Prentice Hall, **(2012)**.
[4] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey", circuits and systems for video technology, ieee transactions on, vol. 18, no. 11, **(2008)**, pp. 1473-1488.
[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, **(2010)**, pp. 1627-1645.
[6] M. Ahmad and L. Seong-Whan, "HMM-based human action recognition using multiview image sequences", 18th International Conference on Pattern Recognition, **(2006)** August 20-24, Hong Kong.
[7] M. Al-Hames and G. Rigoll, "A multi-modal graphical model for robust recognition of group actions in meetings from disturbed videos", IEEE International Conference on Image Processing, **(2005)** September 11-14, Genoa, Italy.
[8] P. Natarajan, V. K. Singh, and R. Nevatia, "Learning 3D action models from a few 2D videos for view invariant action recognition", IEEE Conference on Computer Vision and Pattern Recognition, **(2010)** June 13-18, San Francisco, USA.
[9] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking", 12th European Conference on Computer Vision, **(2012)** October 8-11, Florence, Italy.
[10] D. L. Donoho, "Compressed sensing", IEEE Transactions on Information Theory, vol. 52, no. 4, **(2006)**, pp. 1289 - 1306.
[11] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, **(2012)**, pp. 2274-2282.
[12] M. A. F. R. A. Elschlager, "The representation and matching of pictorial structures", IEEE Transactions on Computers, vol. 22, no. 1, **(1973)**, pp. 67-92.
[13] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark", IEEE Conference on Computer Vision and Pattern Recognition, **(2013)** June 23-27, Portland, USA.

Authors

Gaofeng Li, he received his Master's degree in Physics and Electronics from South China University of Technology in Guangzhou, China. He is a Ph.D. candidate in School of Electronics and Information Engineering, Tongji University, in Shanghai, China. His research interest is mainly in the area of Computer Vision, Pattern Recognition, Detection and Automation.

Fei Wang, he received master degree in Control Science and Engineering from Shanghai University in Shanghai, China, is currently pursuing Ph. D. in School of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests conclude Pattern Recognition, Artificial Intelligent Systems, and Discrete Event Systems.

Wang Lei, he received his Ph.D. degree in Intelligent Control from Gesamthochschule Essen University in German. He is a professor in School of Electronics and Information Engineering, Tongji University, in Shanghai, China. His research interest is mainly in the area of Pattern Recognition, Control and Automation.