HUMAN INTERACTION RECOGNITION USING LOW-RANK MATRIX APPROXIMATION AND SUPER DESCRIPTOR TENSOR DECOMPOSITION

Muhammad Rizwan Khokher, Abdesselam Bouzerdoum, Son Lam Phung

School of Electrical, Computer and Telecommunication Engineering
University of Wollongong, NSW, 2522, Australia
mrk840@uowmail.edu.au, a.bouzerdoum@uow.edu.au, phung@uow.edu.au

ABSTRACT

Audio-visual recognition systems rely on efficient feature extraction. Many spatio-temporal interest point detectors for visual feature extraction are either too sparse, leading to loss of information, or too dense resulting in noisy and redundant information. Furthermore, interest point detectors designed for a controlled environment can be affected by camera motion. In this paper, a salient spatio-temporal interest point detector is proposed based on a low-rank and group-sparse matrix approximation. The detector handles the camera motion through a short-window video stabilization. The multi-modal audio-visual features from multiple descriptors are represented by a super descriptor, from which a compact set of features is extracted through a tensor decomposition and feature selection. This tensor decomposition retains the spatio-temporal structure among features obtained from multiple descriptors. Experimental validation is conducted using two benchmark human interaction recognition datasets: TVHID and *Parliament*. Experimental results are presented which show that the proposed approach outperforms many state-of-the-art methods, achieving classification rates of 74.7% and 88.5% on the TVHID and *Parliament* datasets, respectively.

Index Terms— Human interaction recognition, spatio-temporal interest point detection, low-rank and group-sparse matrix approximation, tensor decomposition

1. INTRODUCTION

Exploiting multi-modal information (e.g., auditory and visual) can greatly improve the performance of recognition systems [1–4]. Audio-visual recognition methods have appeared in different applications, including human interaction recognition [1, 2], action recognition [3] and event detection [4]. A general recognition system, first, extracts multi-modal audio-visual features using multiple descriptors, represents the extracted features using some bag-of-words (BoW) model, and then performs recognition using a classifier. The approach presented here adapts a similar pipe-line, contributing to the first two components: visual feature extraction and multi-modal feature representation.

Most of the visual feature extraction methods, first, detect spatio-temporal interest points (STIPs) and then compute feature descriptors within a volume, either around the STIPs [5] or along trajectories formed by tracking those STIPs [6]. However, there are a few limitations and problems associated with available STIP detectors. Firstly, the detectors are either too sparse, leading to loss of information, or too dense, resulting in irrelevant and redundant information and an increase in computation [7]. A good density of interest points can be realized by the fact that a detector should detect salient interest points representing motion areas only. Secondly, in case of dynamic background and moving camera, the detectors may detect irrelevant interest points that do not belong to actual motion. This is because the detectors are usually designed for a constrained environment where they target local spatio-temporal information without considering global motion [7]. To address these limitations, a STIP detector based on a low-rank and group-sparse (LRGS) matrix approximation is designed, which can consider long-term temporal interactions to extract salient interest points that are not affected by camera motion.

The extracted multi-modal features from multiple feature descriptors are usually represented by the BoW model, where local features are first encoded, then the encoded multi-modal features are concatenated to form a single large feature vector for classification [8, 9]. The concatenation of features from multiple descriptors destroys the spatio-temporal structure among the features. To address this problem, feature descriptors are arranged into a tensor (i.e., multi-dimensional array). This representation provides a natural way to retain the spatio-temporal structure among the features. Then a tensor decomposition is applied to extract salient discriminative features and remove redundant features to achieve dimensionality reduction.

The remainder of the paper is organized as follows. Section 2 describes multi-modal feature extraction and presents a new method for salient spatio-temporal interest point detection. Section 3 presents a super descriptor tensor decomposition model for feature representation. Experimental results and analysis are given in Section 4. Finally, Section 5 con-

cludes the paper.

2. MULTI-MODAL FEATURE EXTRACTION

This section describes the auditory and visual features employed here for human interaction recognition.

2.1. Audio feature extraction

The typical audio feature extraction methods include zero-crossing [10], linear predictive coding [11], and mel-frequency cepstral coefficients (MFCC) [12]. Although, any of these methods can be used in our recognition system, here we employ MFCC features along with their first and second derivatives, which are commonly used in speech recognition. The MFCCs are computed within short audio frames with some overlap.

2.2. Visual feature extraction

To extract visual features, firstly the STIPs are detected using the proposed detector (described in Subsection 2.2.1). Secondly, the detected interest points are tracked up to L frames using optical flow to form motion trajectories. Finally, histogram of oriented gradient (HOG) and motion boundary histogram (MBH) descriptors are computed along the trajectories. The trajectory formation and descriptor computation are adapted from [6].

2.2.1. STIP detector based on LRGS matrix approximation

In this subsection, a new STIP detector based on LRGS matrix approximation is proposed. For a video, spatial interest points (SIPs) are detected in each frame, using FAST corners [13], Canny edges [14], and SURF features [15]. A set of salient STIPs is then detected by considering long-term temporal interactions based on a LRGS matrix approximation.

Each frame is first scanned for the detected SIPs, using row-wise, column-wise or zig-zag scanning. The succeeding $L-1$ frames are then realigned with the scanned frame so as to compensate for global camera motion—we refer to this as short-window video stabilization (described in Subsection 2.2.2). For the ith SIP detected at pixel (x_i, y_i), an L-dimensional column vector $\mathbf{v}_i$ is formed, which contains the pixel values $I(x_i, y_i)$ in the L frames, $\mathbf{v}_i = [I(x_i, y_i, t_j)],\ j = 0, 1, ..., L-1$. Next a matrix $\Phi \in \mathbb{R}^{L \times N}$ is formed using all the SIP vectors $\mathbf{v}_i,\ i = 1, 2, ..., N$, as columns,

$$\Phi = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N]. \tag{1}$$

We consider the decomposition of Φ into low-rank and group-sparse components in presence of noise, as follows:

$$\Phi = B + F + \aleph, \tag{2}$$

where B is a low-rank matrix, F is group-sparse matrix, and $\aleph$ represents additive noise. Based on the low-rank and group-sparsity constraints, the following objective function is to be minimized:

$$\min_{B,F} \|\Phi - B - F\|_F^2 + \lambda \|F\|_{2,1},\ s.t.\ rank(B) \leq R, \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_{2,1}$ is a mixed $\ell_{2,1}$ norm, λ is a regularization parameter to control the sparsity in F, and R represents an upper bound on the rank of B. The optimization problem in (3) is divided into two sub-problems [16] which are solved alternately:

$$B = \min_B \|\Phi - B - F\|_F^2,\ s.t.\ rank(B) \leq R, \tag{4}$$

$$F = \min_F \|\Phi - B - F\|_F^2 + \lambda \|F\|_{2,1}. \tag{5}$$

A greedy alternating minimization approach is used to solve the two sub-problems (4) and (5). By the singular value decomposition of $\Phi - F$ we have, $(\Phi - F) = U\Sigma V^T$, where U and V are matrices of left and right singular vectors, respectively, and Σ is a diagonal matrix of singular values σ_i. The matrix B in (4) can be approximated by the first r dominant singular components,

$$B = \sum_{i=1}^{r} \sigma_i U_i V_i^T, \quad r \leq R. \tag{6}$$

Starting from $r = 1$, the value of r is incremented by 1 at each iteration by checking the contribution of last singular value against the rest, if $(\sigma_r / \sum_{i=1}^{r} \sigma_i)$ is greater than some threshold, then $r = r + 1$; otherwise, the iteration stops.

To solve sub-problem (5), a soft-thresholding based shrinkage operation is applied to approximate F. Let ϕ_i denote the ith column of $\Phi - B$. The ith column of matrix F, F_i is obtained as,

$$F_i = \phi_i \,.\, \max\left(0, 1 - \frac{\lambda}{\|\phi_i\|_2}\right). \tag{7}$$

At each iteration, an error $\|\Phi - B - F\|_F^2 / \|\Phi\|_F^2$ is computed. The alternating minimization terminates if the error becomes less than a threshold ϵ or the number of iterations reaches a maximum set number of iterations. Each column in F corresponds to a SIP. If a column in F is zero, the corresponding SIP belongs to static background, and if the column is non-zero, the SIP belongs to moving foreground. The extracted foreground SIPs form the desired STIPs.

2.2.2. Short-window video stabilization

In the case of camera motion, there can be many unwanted STIPs. To solve this problem, the video frames are spatially aligned on short-window bases by estimating global motion. For a frame t_0, the subsequent frames $t_1 + 1$ to $t_0 + L - 1$ are aligned with it by matching previously extracted SURF points. The SURF descriptors are computed for the SURF points, and the locations of the corresponding points in two frames are retrieved by matching their SURF descriptors. An affine transformation corresponding to the matched point pairs is calculated using M-estimator SAmple Consensus algorithm [17]. Using the estimated geometric transformation, the two frames are aligned. This stabilizes the background within a short-window using the global motion.

3. SUPER DESCRIPTOR TENSOR DECOMPOSITION

In this section, a super descriptor tensor decomposition (SDTD) model is presented to represent multi-modal features from multiple descriptors. The descriptors are encoded and arranged in the form of tensors in order to retain the spatio-temporal structure among the features. The tensors are decomposed using TUCKER-3 decomposition followed by Fisher ranking [18] to obtain salient features for classification.

3.1. Tensor Decomposition

Local audio-visual descriptors, namely MFCC, HOG and MBH, are encoded using super descriptor vector (SDV) coding [19]. The SDV encoded features are then arranged into a $K \times M$ matrix, where K refers to the dictionary size in SDV coding and M is the dimension of the feature vectors. For P different feature descriptors (i.e., MFCC, HOG, MBHx and MBHy), the resultant $K \times M$ matrices are arranged as a rank-3 tensor. For each video segment, a super descriptor tensor of size $K \times M \times P$ is obtained.

The tensor decomposition is employed to discard the noisy and redundant features. Assume a training set of Q rank-3 tensors $\mathbf{X}^i \in \mathbb{R}^{K \times M \times P}$, $i = 1, 2, ..., Q$. The tensor decomposition of $\mathbf{X}^i$ to get three basis factors $A^{(1)} \in \mathbb{R}^{K \times J_1}$, $A^{(2)} \in \mathbb{R}^{M \times J_2}$ and $A^{(3)} \in \mathbb{R}^{P \times J_3}$ and a core feature tensor $\mathbf{G}^i \in \mathbb{R}^{J_1 \times J_2 \times J_3}$ of total $J_1 J_2 J_3$ features, is given as,

$$\mathbf{X}^i \approx \mathbf{G}^i \times_1 A^{(1)} \times_2 A^{(2)} \times_3 A^{(3)}, \tag{8}$$

where $\times_p$, $p = 1, 2, 3$, is the p-mode product of a tensor by a matrix. For example, let $\mathbf{G}^i = \{g_{j_1,j_2,j_3}\}$ and $A^{(1)} = [a_{k,j_1}]$,

$$(\mathbf{G}^i \times_1 A^{(1)})_{k,j_2,j_3} = \sum_{j_1=1}^{J_1} g_{j_1,j_2,j_3}\ a_{k,j_1}, \tag{9}$$

The basis factor $A^{(p)}$ can be obtained by minimizing the following cost function,

$$\min_{\{A^{(1)},A^{(2)},A^{(3)}\}} \sum_{i=1}^{Q} \|\mathbf{X}^i - \mathbf{G}^i \times_1 A^{(1)} \times_2 A^{(2)} \times_3 A^{(3)}\|_F^2. \tag{10}$$

The Q simultaneous standard decompositions of rank-3 tensors $\mathbf{X}^i$ in (8) are equivalent to the following tensor decomposition:

$$\mathbf{X} \approx \mathbf{G} \times_1 A^{(1)} \times_2 A^{(2)} \times_3 A^{(3)}, \tag{11}$$

where the tensors $\mathbf{X} \in \mathbb{R}^{K \times M \times P \times Q}$ and $\mathbf{G} \in \mathbb{R}^{J_1 \times J_2 \times J_3 \times Q}$ are rank-4 tensors obtained by concatenating all the tensors $\mathbf{X}^i$ and $\mathbf{G}^i$ along mode-4. This decomposition model is called the TUCKER-3 tensor decomposition. For a more detailed mathematical description, see [18].

To find the basis factors, higher order orthogonal interactions (HOOI) [20] algorithm is used. The orthogonal interactive basis factors are estimated as $A^{(p)} = U^{(p)}$ of the TUCKER-3 decomposition of the training tensor $\mathbf{X}$. First, the factors $U^{(p)}$ are randomly initialized so that the training core tensor $\mathbf{G}$ can be obtained,

$$\mathbf{G} = \mathbf{X} \times_1 U^{(1)\,T} \times_2 U^{(2)\,T} \times_3 U^{(3)\,T}. \tag{12}$$

Minimizing the cost function in (10) is equivalent to maximizing over the matrices $U^{(p)}$ the function [20],

$$J(U^{(p)}) = \|\mathbf{X} \times_1 U^{(1)\,T} \times_2 U^{(2)\,T} \times_3 U^{(3)\,T}\|_F^2, \tag{13}$$

If $U^{(p)}$ is fixed, the tensor $\mathbf{X}$ can be projected onto the subspace defined as,

$$\begin{aligned} \mathbf{W}^{(-p)} &= \mathbf{X} \times_1 U^{(1)\,T} \times_2 U^{(2)\,T} \times_3 U^{(3)\,T} \\ &= \mathbf{X} \times_{-(p,4)} \{U^T\}, \end{aligned} \tag{14}$$

where $\times_{-(p,4)}$ represents the multiplication excluding mode-p and mode-4. The factors $U^{(p)}$ can be estimated as J_p which are leading left vectors of the mode-p matricized version of $W_{(p)}^{(-p)}$. Once the basis factors $U^{(p)}$ are obtained, a test feature core tensor $\mathbf{G}^t$ for a test tensor $\mathbf{X}^t$ can be obtained as $\mathbf{G}^t = \mathbf{X}^t \times \{U^T\}$.

3.2. Feature Selection and Classification

It is likely that some discriminative features will be lost if the size of the core tensor is set too small during the tensor decomposition. But avoiding feature loss will lead to a large core tensor and inefficient classification. To solve this problem, the salient features for classification are selected using Fisher ranking. The features are sorted in a descending order of their Fisher score. The top features that can achieve the highest classification accuracy are selected through experimentation. For the classification, a one-vs-the-rest linear SVM is used.

4. RESULTS AND ANALYSIS

The proposed STIP detector (i.e., LRGS-STIP) and multi-modal feature representation model (i.e., SDTD) are tested on two publicly available datasets: TV human interaction dataset (TVHID) [21] and *Parliament* [22]. For the sake of fair comparison, we adopt the same evaluation protocols as in [21] and [22]; that is, ten-fold and five-fold cross-validation are employed with datasets TVHID and *Parliament*, respectively. We set $R = 15$ and $\epsilon = 0.001$. These values work for both datasets. The parameter λ is tuned separately for each dataset: it is set to $\lambda = 0.03$ for TVHID and $\lambda = 0.05$ for *Parliament*. The SIPs are tracked up to $L = 15$ frames. For the trajectory and descriptor computation, the same settings are used as in [6]. From each trajectory and audio frame, a ninety-six ($M = 96$) dimensional feature vector is calculated for each descriptor (i.e., MFCC, HOG, MBHx and MBHy). For SDV coding, the dictionary size is set to 500, which is suitable for a wide range of datasets [19]. Therefore, a tensor of size $500 \times 96 \times 4$ is obtained for each video sample. We test different tensor decomposition algorithms (i.e., higher order discriminant analysis (HODA) and HOOI) and feature ranking criteria (i.e., Student's t-test, mutual information and Fisher ranking). From experiments, the TUCKER-3 decomposition with HOOI algorithm and Fisher feature ranking give the best classification accuracy.

In the first experiment, performance of the LRGS-STIP detector is compared with Harris3D [5] and dense sampling

[6] methods. For the three methods, the trajectories and descriptors are computed from the STIPs found by their respective detectors. The descriptors are encoded using SDV, and the encoded features are concatenated, to obtain a single vector for classification. The audio features and tensor decomposition are not used in this experiment. The classification rates (CR) of the LRGS-STIP, Harris3D and dense sampling methods are presented in Table 1. The LRGS-STIP achieves average CRs of 67.3% on TVHID and 83.7% on *Parliament*; it outperforms both the Harris3D and dense sampling methods. The Harris3D is a sparse detector, which may have resulted in loss of information; it achieves average CRs of 41.8% and 73.8%. The dense sampling results in too many STIPs which increases the number of noisy features and computation. Although the dense sampling achieves higher CRs than the Harris3D detector (53.4% and 79.2%), its CRs are still much lower than those of LRGS-STIP. The Harris3D and dense sampling do not take into account the camera motion and result in many irrelevant STIPs which may have affected their CRs. On the other hand, the LRGS-STIP considers long-term temporal interactions and extracts a salient set of STIPs belonging to actual motion, and hence achieves the highest CRs.

Table 1: Average CRs and standard error in % of the proposed LRGS-STIP, Harris3D [5] and dense sampling [6] detectors.

	LRGS-STIP	**Harris3D**	**Dense Sampling**
TVHID	**67.3 ± 0.3**	41.8 ± 1.2	53.4 ± 0.6
Parliament	**83.7 ± 0.2**	73.8 ± 0.6	79.2 ± 0.3

In the second experiment, the proposed SDTD model is compared with three other feature representation methods: LLC [8], FV [9] and SDV [19]. Both audio and visual features are used in this experiment. For the LLC, FV and SDV-1, the dictionary size is set to 500, and the feature descriptors are concatenated to form a large single vector for classification. For the SDTD model, a salient set of features is extracted after the tensor decomposition and Fisher ranking. For the SDV-2, the same number of features are selected as in SDTD, by applying only Fisher ranking without tensor decomposition. The total number of features f used for classification and the classification results of each method are shown in Table 2. Among the existing methods, SDV-1 outperforms LLC, FV and SDV-2 on both TVHID and *Parliament* datasets. However, the proposed SDTD method outperforms all other methods; it achieves the highest CRs of 74.7% and 88.5% on the two datasets. The SDTD retains the spatio-temporal structure among features from multiple descriptors, which is usually destroyed if the features are concatenated. The tensor decomposition discards the noisy and redundant features which can affect the classifier accuracy. Although the CR of SDTD is only slightly higher than that of SDV-1, a significant reduction in the feature dimension is achieved by the SDTD model: the number of features are reduced from 192000 to 3500 and 2500 for the two datasets.

In the last experiment, the proposed pipe-line (i.e, LRGS-STIP + SDTD) is compared with some other methods for human interaction recognition task. The other methods use different type of features (i.e., audio-visual and visual only). The CRs of the proposed approach and other methods are given in Table 3. The CRs of the other methods are directly taken from the references shown in the table. The LRGS-STIP + SDTD approach outperforms most of the other methods on the TVHID and *Parliament* datasets, when using audio-visual and visual only features. Only the SAVAR [2] method, which extracts more features such as head orientation and proxemic, outperforms our proposed method.

Table 2: Average CRs and standard error in % of the SDTD method and different feature representation methods (i.e., LLC [8], FV [9] and SDV [19]).

	SDTD ($f = 3500, 2500$)	**LLC** ($f = 2000$)	**FV** ($f = 192000$)	**SDV-1** ($f = 192000$)	**SDV-2** ($f = 3500, 2500$)
TVHID	**74.7 ± 0.5**	72.3 ± 1.1	73.0 ± 0.9	73.7 ± 0.8	69.2 ± 1.3
Parliament	**88.5 ± 0.3**	85.3 ± 0.8	86.5 ± 0.5	87.2 ± 0.6	80.0 ± 0.9

Table 3: Average CRs in % of the proposed LRGS-STIP + SDTD and other methods.

	TVHID		***Parliament***	
Methods	**CR**	**Features type**	**CR**	**Features type**
Patron *et al.* [21]	54.7	Visual	NA	NA
Yu *et al.* [23]	66.2	Visual	NA	NA
Li *et al.* [24]	68.0	Visual	NA	NA
Marin *et al.* [1]	54.5	Audio-visual	NA	NA
SAVAR [2]	**81.3**	Audio-visual	**97.6**	Audio-visual
Vrigkas *et al.* [22]	NA	NA	85.5	Visual
LRGS-STIP + SDTD	69.1	Visual	83.7	Visual
LRGS-STIP + SDTD	74.7	Audio-visual	88.5	Audio-visual

5. CONCLUSION

This paper presents a new spatio-temporal interest point detector based on a low-rank and group-sparse matrix approximation. The proposed detector extracts a salient set of interest points by taking into account long-term temporal interactions and camera motion. The experiments show that the proposed detector outperforms some existing detectors. The multi-modal audio-visual features from multiple descriptors are represented by a super descriptor tensor decomposition model, in order to retain spatio-temporal structure among features and obtain more discriminative features for classification. The tensor decomposition followed by Fisher ranking discards the noisy and redundant features. In comparison with other feature representation methods, the proposed tensor decomposition model achieves a significant reduction in features along with a higher classification rate. Furthermore, the overall proposed audio-visual recognition system outperforms many existing methods on the same task of human interaction recognition.

Acknowledgment

This work has been supported in part by a grant from the Australian Research Council.

6. REFERENCES

[1] M. J. Marin-Jimenez, R. M. noz Salinas, E. Yeguas-Bolivar and N. P. de la Blanca, "Human interaction categorization by using audio-visual cues," *MVA*, vol. 25, no. 1, pp. 71–84, 2014.

[2] M. Vrigkas, C. Nikou and I. Kakadiaris, "Identifying Human Behaviors Using Synchronized Audio-Visual Cues," *AC*, 2015. doi: 10.1109/TAFFC.2015.2507168

[3] Q. Wu, Z. Wang, F. Deng, Z. Chi and D. D. Feng, "Realistic human action recognition with multimodal feature selection and fusion," *SMC*, vol. 43, no. 4, pp. 875–885, 2013.

[4] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," *In Proc. ICASSP*, pp. 3553–3556, 2009.

[5] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *In Proc. VS-PETS*, pp. 65–72, 2005.

[6] H. Wang, A. Klaser, C. Schmid and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, pp. 60–79, 2013.

[7] D. D. Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," *VC*, vol. 32, no. 3, pp. 289–306, 2016.

[8] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang and Y. Gong, "Locality constrained linear coding for image classification," *In Proc. CVPR*, pp. 3360–3367, 2010.

[9] J. Sanchez, F. Perronnin, T. Mensink and J. Verbeek, "Image classification with the fisher vector: theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, 2013.

[10] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *IEEE Proceedings*, vol. 74, pp. 1477–1493, 1986.

[11] L. Rabiner and R. Schafer, "Digital Processing of Speech Signals," Prentice Hall Inc., Englewood Cliffs, New Jersey, 1978.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *ASSP*, vol. 28, no. 4, pp. 357–366, 1980.

[13] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *In Proc. ECCV*, pp. 430–443, 2006.

[14] J. Canny, "A computational approach to edge detection," *PAMI*, vol. 8, no. 6, pp. 679–698, 1986.

[15] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, "SURF: Speeded up robust features," *CVIU*, vol. 110, no. 3, pp. 346–359, 2008.

[16] T. Zhou and D. Tao, "Shifted subspaces tracking on sparse outlier for motion segmentation," *In Proc. IJCAI*, pp. 1946–1952, 2013.

[17] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *CVIU*, vol. 78, pp. 138–156, 2000.

[18] A. H. Phan and A. Cichocki, "Tensor decompositions for feature extraction and classification of high dimensional datasets," *NTA, IEICE*, vol. 1, no. 1, pp. 37–68, 2010.

[19] X. Yang and Y. Tian, "Action recognition using super sparse coding vector with spatio-temporal awareness," *In Proc. ECCV*, pp. 727–741, 2014.

[20] L. D. Lathauwer, B. D. Moor and J. Vandewalle, "On the best rank-1 and rank-(R1,R2, . . . ,RN) approximation of higher-order tensors," *SIAM JMAA*, vol. 21, no. 4, pp. 1324–1342, 2000.

[21] A. Patron-Perez, M. Marszalek, I. Reid and A. Zisserman, "Structured learning of human interactions in TV shows," *PAMI*, vol. 34, no. 12, pp. 2441–2453, 2012.

[22] M. Vrigkas, C. Nikou and I. A. Kakadiaris, "Classifying behavioral attributes using conditional random fields," *In Proc. HCAI*, pp. 95–104, 2014.

[23] G. Yu, J. Yuan and Z. Liu, "Propagative Hough voting for human activity recognition," *In Proc. ECCV*, pp. 693–706, 2012.

[24] B. Li, M. Ayazoglu, T. Mao, O. I. Camps and M. Sznaier, "Activity recognition using dynamic subspace angles," *In Proc. CVPR*, pp. 3193–3200, 2011.