Human Action Attribute Learning Using Low-Rank Representations

Tong Wu*, Prudhvi Gurram†‡, Raghuveer M. Rao‡ and Waheed U. Bajwa*
*Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854
†Booz Allen Hamilton, McLean, VA 22102 ‡U.S. Army Research Laboratory, Adelphi, MD 20783

***Abstract*—This paper studies the problem of learning human action attributes based on union-of-subspaces model. It puts forth an extension of the low-rank representation (LRR) model, termed the hierarchical clustering-aware structure-constrained low-rank representation (HCS-LRR) model, for unsupervised learning of human action attributes from video data. The effectiveness of the proposed model is demonstrated through experiments on five human action datasets for action recognition.**

I. Introduction

Human activities comprising a sequence of actions can be represented hierarchically [1]. The bottom level of this hierarchy contains multiple action attributes, which describe an action at the finest resolution [2]. In this paper, we propose to represent human action attributes based on the *union-of-subspaces* (UoS) model [3], where each action attribute corresponds to one of the subspaces. A human action or activity can then be uniquely represented as a sequence of transitions from one action attribute to another, which in turn can be used for human action recognition. We propose a *hierarchical clustering-aware structure-constrained LRR* (HCS-LRR) model, for unsupervised learning of action attributes from video data without the need to specify the number of attributes a priori. Our results confirm the superiority of HCS-LRR over the state-of-the-art subspace clustering approaches for action recognition.

II. Hierarchical Clustering-Aware Structure-Constrained Low-Rank Representation

In this work, we learn action attributes using two local visual descriptors: HOG (histograms of oriented gradients) [4] and MBH (motion boundary histogram) [5]. The extracted features are represented by a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{m\times N}$. Each column $\mathbf{x}_i$ has unit ℓ_2 norm and corresponds to the feature vector of a frame and n_τ adjacent optical flow fields for HOG and MBH features, respectively. Suppose these N feature vectors are drawn from a union of L low-dimensional subspaces $\{\mathcal{S}_\ell\}_{\ell=1}^L$ of dimensions $\{d_\ell\}_{\ell=1}^L$. The *clustering-aware structure-constrained LRR* (CS-LRR) model amounts to solving the following optimization problem [6]:

$$\min_{\mathbf{Z},\mathbf{F},\mathbf{E}} \|\mathbf{Z}\|_* + \alpha\|\mathbf{B}\odot\mathbf{Z}\|_1 + \beta\mathrm{tr}\Big(\mathbf{F}^T(\mathbf{V} - \frac{|\mathbf{Z}|+|\mathbf{Z}^T|}{2})\mathbf{F}\Big) + \lambda\|\mathbf{E}\|_\iota$$
$$\text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E},\ \mathbf{F}^T\mathbf{F} = \mathbf{I}, \tag{1}$$

where $\odot$ denotes the Hadamard product and $\|\cdot\|_\iota$ indicates a certain regularization strategy involving $\mathbf{E}$. The (i,j)-th entry of $\mathbf{B}$ is defined as $b_{i,j} = 1 - \exp(-\frac{1-|\mathbf{x}_i^T\mathbf{x}_j|}{\sigma})$, where σ is the mean of all $1 - |\mathbf{x}_i^T\mathbf{x}_j|$'s, while $\mathbf{F} \in \mathbb{R}^{N\times\mathcal{L}}$ is a binary matrix indicating the cluster membership of the data points. Defining an affinity matrix $\mathbf{W}$ as $\mathbf{W} = \frac{|\mathbf{Z}|+|\mathbf{Z}^T|}{2}$, the matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose diagonal elements are defined as $v_{i,i} = \sum_j w_{i,j}$. Problem (1) can be solved by using the linearized alternating direction method [7].

Once we have obtained the optimal representation coefficient matrix $\widehat{\mathbf{Z}}$ by solving Problem (1), we set the coefficients below a given threshold to zeros and we denote the final representation matrix by $\widetilde{\mathbf{Z}}$. We now define the affinity matrix as $\mathbf{W} = \frac{|\widetilde{\mathbf{Z}}|+|\widetilde{\mathbf{Z}}^T|}{2}$ and proceed with our hierarchical clustering procedure as follows. Let $\mathbf{X}_{p|\ell}$ be the collection of samples belonging to the ℓ-th cluster at the p-th level ($p \geq 0$) and $\boldsymbol{\pi}_{p|\ell}$ denote the set containing the indices of all $\mathbf{x}_i$'s that are assigned to $\mathbf{X}_{p|\ell}$. We start at the top with all data points in one cluster, i.e., $\mathbf{X}_{0|0} = \mathbf{X}$ and $\boldsymbol{\pi}_{0|0} = \{1, 2, \ldots, N\}$. Then at each level $p \geq 0$, we split $\mathbf{X}_{p|\ell}$ into two sub-clusters by applying spectral clustering [8] on $[\mathbf{W}]_{\boldsymbol{\pi}_{p|\ell},\boldsymbol{\pi}_{p|\ell}}$ (the submatrix of $\mathbf{W}$ whose rows and columns are indexed by $\boldsymbol{\pi}_{p|\ell}$). Note that when $p \geq 2$, we have an additional step which decides whether or not to further divide each single cluster (i.e., subspace) at the p-th level into two clusters (subspaces) at the $(p+1)$-th level. The cluster $\mathbf{X}_{p|\ell}$ is divisible if and only if (i) the relative representation errors of the data samples using the child subspace are less than the representation errors calculated using the parent subspace by a certain threshold, and (ii) the dimensions of the two child subspaces meet a minimum requirement. Otherwise the cluster $\mathbf{X}_{p|\ell}$ is a leaf cluster and this cluster will not be divided any further. This process is repeated until we reach a predefined maximum level in the hierarchy denoted by P. We term our hierarchical subspace clustering algorithm based on CS-LRR model as HCS-LRR. Note that the maximum number of leaf clusters is $L_{\max} = 2^P$ in this setting, which we set as a key input parameter in Problem (1). We omit further details here and refer the reader to [9] for a full explanation of the algorithm.

III. Experiments

We now evaluate our proposed method for human action recognition on five datasets: Weizmann [10], Ballet [11], UIUC [12], Keck [13] and UCF Sports [14]. We compare the performance of HCS-LRR with LRR [3], SSC [15], SC-LRR [16], and LSR [17]. The number of clusters L for these algorithms is set (i) to be the same number of leaf subspaces generated by HCS-LRR (denoted by ⟨*Algorithm*⟩-L_P) and (ii) to be the same as the number of actions in the training data (denoted by ⟨*Algorithm*⟩-Q). For classification we use a non-linear SVM [18] with a Gaussian Dynamic Time Warping (DTW) kernel [19], where the distance between two video sequences is computed by applying DTW [20] on the Grassmann manifold. This only involves subspace transition vectors of the videos. We use both *one-vs.-all* and *one-vs.-one* approach for classification, denoted by "SVM/ova" and "SVM/ovo" in Table I, respectively. The classification results are listed in Table I, from which we make the conclusion that by representing the human actions using the attributes learned by HCS-LRR, we are always able to recognize the actions at a higher rate compared to other subspace clustering methods.

This work is supported in part by the ARL, ARO, and NSF.

TABLE I
ACTION RECOGNITION RESULTS (%)

Dataset	Feature	Classifier	Subspace clustering method								
			HCS-LRR	LRR-L_P	LRR-Q	SSC-L_P	SSC-Q	SC-LRR-L_P	SC-LRR-Q	LSR-L_P	LSR-Q
Weizmann	HOG	SVM/ova	**92.22**	65.56	52.22	58.89	48.89	64.44	63.33	58.89	65.56
		SVM/ovo	**95.56**	82.22	66.67	64.44	51.11	68.89	68.89	71.11	72.22
	MBH	SVM/ova	**87.78**	85.56	64.44	85.56	64.44	81.11	68.89	86.67	71.11
		SVM/ovo	85.56	**87.78**	66.67	84.44	61.11	81.11	66.67	85.56	71.11
Ballet	HOG	SVM/ova	**67.80**	59.32	30.51	**67.80**	38.98	66.10	42.37	54.24	61.02
		SVM/ovo	62.71	62.71	52.54	**64.41**	42.37	**64.41**	57.63	54.24	61.02
	MBH	SVM/ova	**71.19**	69.49	61.02	54.24	20.34	61.02	62.71	67.80	67.80
		SVM/ovo	**69.49**	67.80	61.02	45.76	44.07	67.80	61.02	62.71	67.80
UIUC	HOG	SVM/ova	**100**	77.14	91.43	98.57	81.43	**100**	90.00	**100**	91.43
		SVM/ovo	**100**	92.86	98.57	**100**	78.57	**100**	91.43	**100**	85.71
	MBH	SVM/ova	**100**	**100**	95.71	**100**	**100**	**100**	78.57	**100**	**100**
		SVM/ovo	**100**	**100**	95.71	**100**	**100**	**100**	77.14	**100**	**100**
Keck	MBH	SVM/ova	**87.30**	76.98	66.67	74.60	71.43	79.37	80.16	57.14	61.11
		SVM/ovo	**90.48**	76.98	74.60	78.57	73.81	84.92	84.13	70.63	69.05
UCF	MBH	SVM/ova	66.13	**67.74**	60.48	66.13	54.84	64.52	50.81	64.52	54.03
		SVM/ovo	**75.81**	75.00	57.26	**75.81**	53.23	73.39	45.16	68.55	50.00

REFERENCES

[1] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimed. Inf. Retr.*, vol. 2, no. 2, pp. 73–101, 2013.

[2] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 3337–3344.

[3] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, 2013.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[5] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 428–441.

[6] T. Wu, P. Gurram, R. M. Rao, and W. U. Bajwa, "Clustering-aware structure-constrained low-rank representation model for learning human action attributes," in *Proc. IEEE Image Video and Multidimensional Signal Process. (IVMSP) Workshop*, 2016, pp. 1–5.

[7] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 612–620.

[8] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 849–856.

[9] T. Wu, P. Gurram, R. M. Rao, and W. U. Bajwa, "Human action attribute learning from video data using low-rank representations," *arXiv:1612.07857*, 2016.

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[11] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[12] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 548–561.

[13] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 444–451.

[14] M. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[15] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[16] K. Tang, R. Liu, Z. Su, and J. Zhang, "Structure-constrained low-rank representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2167–2179, 2014.

[17] C. Lu, H. Min, Z. Zhao, L. Zhu, D. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 347–360.

[18] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[19] S. Gudmundsson, T. P. Runarsson, and S. Sigurdsson, "Support vector machines and dynamic time warping for time series," in *Proc. IEEE Int. Joint Conf. Neu. Net. (IJCNN)*, 2008, pp. 2772–2776.

[20] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.