A HETEROGENEOUS DICTIONARY MODEL FOR REPRESENTATION AND RECOGNITION OF HUMAN ACTIONS

Rushil Anirudh[⋆†], *Karthikeyan Ramamurthy*[†], *Jayaraman J. Thiagarajan*[†], *Pavan Turaga*[⋆†], *Andreas Spanias*[†]

[⋆] School of Arts, Media & Engineering,
[†]School of Electrical, Computer, & Energy Engineering,
Arizona State University, Tempe, AZ.

ABSTRACT

In this paper, we consider low-dimensional and sparse representation models for human actions, that are consistent with how actions evolve in high-dimensional feature spaces. We first show that human actions can be well approximated by piecewise linear structures in the feature space. Based on this, we propose a new dictionary model that considers each atom in the dictionary to be an affine subspace defined by a point and a corresponding line. When compared to centered clustering approaches such as K-means, we show that the proposed dictionary is a better generative model for human actions. Furthermore, we demonstrate the utility of this model in efficient representation and recognition of human activities that are not available in the training set.

Index Terms— Dictionary learning, Sparse representations, Activity analysis.

1. INTRODUCTION

Sparse coding attempts to represent data vectors using a linear combination of a small number of vectors chosen from a 'dictionary'. The dictionary that leads to an optimal sparse representation can be either predefined or learned from the training samples themselves. It is now well known that the latter can lead to improved representation and recognition results [1, 2]. If the data is truly low-dimensional, sparse coding can effectively identify its low degrees of freedom, and hence sparse models have proved successful in several inverse problems in signal/image processing [1], and computer vision [3]. When compared to classical subspace methods which are efficient only if the data lies in a single low-dimensional subspace, sparse coding can recover data lying in a union of low-dimensional subspaces and hence provide a greater flexibility in representation. Traditionally, most sparse coding applications deal with static data such as images, but there have been recent attempts to extend these concepts to videos [4, 5]. To this end, problems of activity analysis have gained lot of attention where typically a dictionary is learned either per class of actions or on the entire set of all actions and sparse codes are generated per frame. Most human actions evolve over time where they usually begin with a rest pose and end in an extreme pose. This transition is smooth resulting in smoothly varying features. The geometric structure of these transitions is not known in general, but attempts have been made to model this structure, e.g. actions have been considered to trace out non-linear manifolds in feature spaces [6]. While such models are quite rich and general, they are accompanied by difficulties in learning the model and coding data using the model. However, as shown in fig 1, a simple piecewise linear model is sufficient to represent most common activities such as *Waving, Running* and *Talking on the phone*. In addition to the representational simplicity, this also affords solving the sparse-coding problem efficiently.

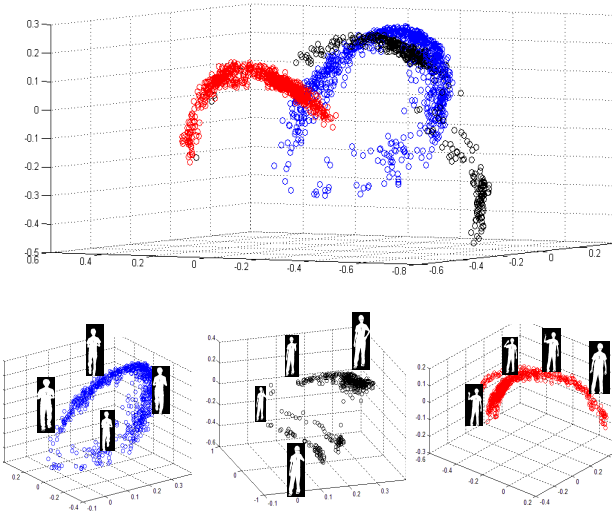

Fig. 1: Here we show the feature evolution of *Running, Talk on Phone* and *Waving*. The features are projected to a lower dimensional space for visualization. The top figure shows the three actions on a common coordinate frame. It is seen that these structures can be well approximated by piece-wise linear models.

In such cases, centered clustering approaches such as K-Means will not be able to effectively model the underlying

Corresponding author - ranirudh@asu.edu (Rushil Anirudh)

patterns which will result in a loss in performance. To cluster data that lies along hyperlines, He *et al.* [7] proposed the K-hyperline clustering algorithm, which is an iterative procedure that performs a least squares fit of K one dimensional linear subspaces to the training data. The relation between K-hyperline clustering and dictionary learning has been explored in [8]. Taking into consideration that cluster centers computed by this algorithm are constrained to pass through the origin, we propose a new heterogeneous dictionary model in this paper. The elementary features in this dictionary correspond to the $1D$ affine subspaces that represent human activities and hence the dictionary is interpretable. The proposed dictionary is learned with features that are extracted per frame from the videos in an action dataset.

Although several dictionary learning approaches are known, only a few have been proposed that consider the geometric structure along which activities evolve. Most of the methods involve improving an initial dictionary, obtained using methods such as K-SVD [1], by maximizing information between dictionary atoms [5], learning class specific dictionaries [4] etc. The idea of features lying along lines has been used before - Taheri *et al.* [9] modeled facial expressions as deviations along geodesics, which are generalizations of high dimensional lines to non Euclidean spaces, from a "neutral expression", and Troje [10] showed that using simple PCA one can identify important directions in landmark data, that are later used for applications like gender classification.

In this paper, we present a dictionary model for human activities by considering piecewise linear models of activities. Each dictionary atom consists of a tuple - a point and a direction in space. We also introduce new constraints to the traditional sparse coding problem, and adapt it to the heterogeneous dictionary. We show that this can be an effective generative model for human actions. Furthermore, we demonstrate that using such a dictionary, one can achieve state-of-the-art recognition results, and maintain very low reconstruction errors for unseen test activities.

2. PROPOSED DICTIONARY MODEL

In this section, we will formulate our dictionary learning problem and present a method to generate sparse codes using the proposed dictionary.

2.1. Learning the Dictionary

When a dictionary is constructed using K-hyperline clustering, each atom corresponds to a linear subspace. In this paper, we generalize this dictionary to be a collection of affine subspaces, where each atom is described by a point and an associated direction in space. To learn such a dictionary, we propose a $1D$ affine subspace clustering algorithm. In this method, we incorporate an additional step of calculating the sample mean $\boldsymbol{\mu}_j$ of the j^{th} cluster along with the least-squares fit of a $1D$ subspace, $\mathbf{d}_j$, in K-hyperline clustering. The algorithm is described in Table 1. To identify the cluster membership, we project a data sample onto each dictionary atom and choose the one that results in the least representation error. The projection is performed as

Input
Features $\{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ and size of dictionary, K.

Output
Affine subspaces $\{\mathbf{H_1}, \cdots, \mathbf{H_K}\}$ represented using the means $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$ and the directions $\{\mathbf{d}_1, \cdots, \mathbf{d}_K\}$.
Membership classes, $C_1, \cdots, C_K$.

Algorithm
Initialize: $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$ and $\{\mathbf{d}_1, \cdots, \mathbf{d}_K\}$.

while convergence not reached
Compute memberships:
- For each sample $\mathbf{x}_i$ compute the projection of $\mathbf{x}_i$ onto each $\mathbf{H}_j$, denoted by $P_{\mathbf{H_j}}(\mathbf{x}_i)$.
- $k = \text{argmin}_j \, ||\mathbf{x}_i - P_{\mathbf{H_j}}(\mathbf{x}_i)||_{j=1}^K$ and $C_k = C_k \cup \{i\}$.

Update $\mathbf{H}_j$: For each cluster j, compute $\{\boldsymbol{\mu}_j, \mathbf{d}_j\}$ as the sample mean and the first principal component of all samples indexed by C_j, respectively.
end

Table 1: The dictionary learning algorithm.

$$P_{\mathbf{H}}(\mathbf{x}) = \boldsymbol{\mu} + \hat{\beta}\mathbf{d}, \quad \text{where} \quad \hat{\beta} = \min_{\beta} \|\mathbf{x} - \boldsymbol{\mu} - \beta\mathbf{d}\|_2^2. \quad (1)$$

Note that in this case, the least squares solution for β is $\mathbf{d}^T(\mathbf{x} - \boldsymbol{\mu})$.

2.2. Sparse Coding

Let us assume that a test sample in $\mathbb{R}^n$ can be represented as a linear combination of a small number of affine subspaces. Assuming that the set of dictionary atoms given by $\{\boldsymbol{\mu}_j, \mathbf{d}_j\}_{j=1}^K$ is known, the generative model for a test sample $\mathbf{x}$ can be written as

$$\mathbf{x} = \sum_{j \in S} \alpha_j \boldsymbol{\mu}_j + \beta_j \mathbf{d}_j. \quad (2)$$

where S is the set of atoms that participate in the representation of $\mathbf{x}$.

The solution to (2) can be obtained using convex programming. The key consideration is that for a given j, $\boldsymbol{\mu}_j$ and $\mathbf{d}_j$ must be chosen together. Furthermore, it is also useful to ensure that the new mean is in the convex hull of the means of S. This can be posed and solved as group Lasso [11],

$$\begin{aligned} \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\text{argmin}} \, \|\mathbf{x} - (\mathbf{M}\boldsymbol{\alpha} + \mathbf{D}\boldsymbol{\beta})\|_2^2 + \lambda \sum_{i=1}^{K} \left\| \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \right\|_2 \\ \text{s.t. } \alpha_i \geq 0, \sum_i \alpha_i = 1, \end{aligned} \quad (3)$$

where $\mathbf{M} = [\boldsymbol{\mu}_j]_{j=1}^K$ and $\mathbf{D} = [\mathbf{d}_j]_{j=1}^K$.

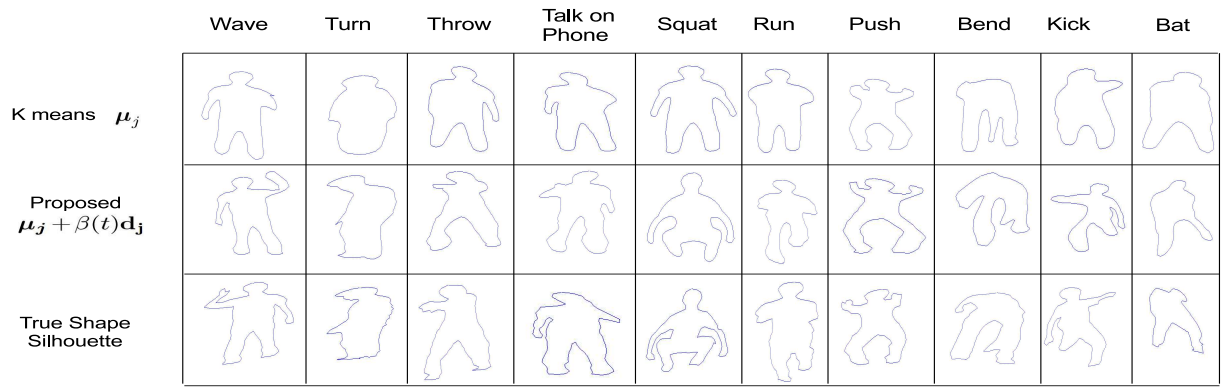

Fig. 2: Actions generated by sampling along the learned lines on the UMD actions data set [12]. Some generated actions such as *wave, talk on phone, kick* appear to be laterally inverted as our representation is affine invariant.

3. EXPERIMENTAL VALIDATION

In this section, we demonstrate the use of the dictionary model in representation and recognition of human actions. First, we perform an experiment to validate the proposed generative model, in comparison to a centered clustering approach. Following this, we show that this dictionary can generalize well in representing unseen human actions. Finally, we demonstrate that by aggregating the sparse codes in multiple temporal scales, we can achieve the state-of-the-art performance in activity recognition.

Generative Model for Human Actions: In this experiment we show that the proposed dictionary can be used to parameterize human actions, thereby demonstrating that the model is an intuitive choice. We perform this experiment using a shape feature due to its obvious advantage in visualization. We use the UMD Actions Dataset [12], as its background is relatively static and allows us to do easy background subtraction. Having extracted the foreground, we perform morphological operations and extract the contour of the human. We sampled a fixed number of points on the contour to obtain the set of landmarks describing the shape. To represent these landmarks, we used an affine invariant representation where the set of m landmark points are given by the $m \times 2$ matrix $L = [(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)]$ for the centered shape. However, shape features do not lie in the Euclidean space [13] and one must take into account the non-linearity of the space while dealing with them. Since in this paper we are dealing with the vector space, we will use embedding approaches as they are conceptually simpler and easier to implement. These allow us to work with these complex features while staying in a Euclidean space. With each set of landmarks, we generate an $m \times m$ projection matrix that is $P = UU^T$, where $L = USV^T$ is the rank-2 SVD. Let P_v be the vectorized form of P, we use P_v as a feature to learn our dictionary. To recover the shape from this vector we re-obtain the projection matrix P and perform a rank-2 SVD on it. Now the feature corresponding to a shape at time t is generated as $P_v(t) = \boldsymbol{\mu_j} + \beta(t)\mathbf{d_j}$, parameterized by $\beta(t)$ which determines to what extent one must travel from $\boldsymbol{\mu_j}$ along the direction $\mathbf{d_j}$. We used different values of β for each action in the range $-1 < \beta(t) < 1$. In fig 2, we show the generated silhouette in each action and compare it to the ground truth.

Reconstruction of Unseen Actions: In this experiment, we test the efficiency of the proposed dictionary in modeling unseen actions from test data. Since every action is modeled as a combination of means and directions, an unseen action will typically have a mean that is different from any of the previously learned actions. Hence, we model the new mean as a linear combination of means and find its principal direction as a combination of the known directions. For our experiments, we obtained activities from the Weizmann activity dataset [14] which consists of 90 videos of 10 different actions, each performed by 9 different persons. The classes of actions include running, jumping, walking, side walking etc. In order to evaluate the performance of the proposed sparse coding model, we used the features of all subjects from 6 different activities in the Weizmann dataset for obtaining the dictionary and evaluated the reconstruction error for features from the other 4 activities. The set of *unseen* testing activities included *jack, pjump, skip* and *wave1*. For all our experiments on this dataset we used the histogram of oriented optical flow (HOOF) feature that was introduced in [15]. This feature bins optical flow vectors based on their directions and their primary angle with the horizontal axis, weighted by their magnitudes. Using magnitudes alone is susceptible to noise and can be very sensitive to scale. Thus all optical flow vectors, $v = [x, y]^T$ with direction $\theta = \tan^{-1}(\frac{y}{x})$ in the range $-\frac{\pi}{2} + \pi\frac{b-1}{B} \leq \theta < -\frac{\pi}{2} + \pi\frac{b}{B}$ will contribute by $\sqrt{x^2 + y^2}$ to the sum in bin b, where $1 \leq b \leq B$, typically $B = 30$ is used. Finally, the histogram is normalized to sum up to 1.

Using the training activities, we computed K (fixed at $20, 30$ and 40) clusters to identify the principal directions and

Method	No. of clusters		
	K=20	K=30	K=40
K-means - $\boldsymbol{\mu}$	0.3295	0.3069	0.2985
K-Hyperline $\mathbf{d}$	0.2657	0.2485	0.2399
$(\boldsymbol{\mu}, \mathbf{d})$ Dictionary	**0.1171**	**0.1039**	**0.0956**

Table 2: Comparison of reconstruction error obtained using the proposed sparse coding with K-means and K-hyperline clustering algorithms. This demonstrates that by using a linear combination of a few known atoms, we can model even unknown actions effectively.

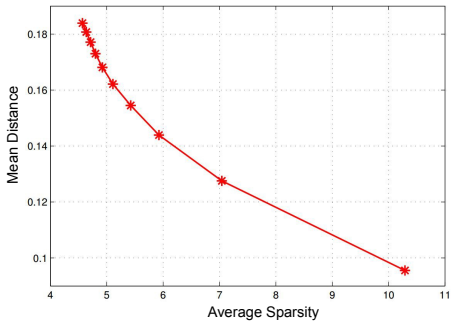

Fig. 3: Effect of sparsity on reconstruction error.

Proposed dictionary	**98.88**
K-means dictionary	84.44
Guha *et al.*, Multiple Dictionaries [4]	**98.9**
Guha *et al.*, Single Dictionary [4]	**96.67**
Chaudry *et al.* [15]	95.66

Table 3: Recognition performance (%) using the sparse codes generated with our model matched the best results on the Weizmann dataset, outperforming several other techniques. As a baseline, the recognition on sparse codes obtained using a K-means dictionary is also shown.

their cluster centroids. For the test activities, we performed sparse coding of the features using the computed centers and directions as the dictionary atoms. Table 2 compares the average reconstruction error obtained for features from the test activities using different coding schemes. Since more than one atom can be used for representation, the reconstruction error in our model is significantly lower than those obtained with K-means or K-hyperline clustering. The plot in Fig 3 shows the reconstruction error obtained by varying the sparsity parameter λ.

Recognition of Human Activities: In this experiment, we propose a method for performing recognition of human activities from the Weizmann dataset using sparse codes obtained from the features of each activity. Of the 9 subjects that performed the activities, we used 6 subjects from each class for training and the rest for testing. Hence, we used a total of 60 activities for learning the dictionary and training the classifier. Using the features described in the previous experiment, the sparse codes are computed by setting $\lambda = 0.1$. We aggregate the sparse codes of the training features, in multiple temporal scales, to create one overall feature vector per activity. Given a set of sparse codes stacked in a matrix, aggregation is performed by finding the value corresponding to the absolute maximum of elements in each row. Since aggregation destroys temporal information, we divide each activity into 1, 2, 4, and 6 temporal segments, and perform aggregation independently in each, in order to partially preserve the temporal information. Hence, if each sparse code is of length K, we will obtain a overall feature vector of length $13K$. These overall feature vectors are used to train an SVM classifier. For a test activity, the overall feature vector is computed similarly and classification is performed.

In order to improve the reliability of recognition results, we repeat the experiment 3 times with randomly chosen training and test sets. Table 3 compares our average performance to other methods reported in the literature. It can be seen that our method compares well with Guha *et al.*, where we are able to match their performance with just a single dictionary as compared to learning a dictionary per class.

4. CONCLUSION AND FUTURE WORK

The proposed model opens up several interesting avenues of research, we outline a few of them and conclude our work in this section.

We introduced a sparse representational model for human actions. We first showed that in feature spaces, common actions are approximately piecewise linear. Using this idea, we proposed a dictionary model where each atom is a $1D$ affine subspace described by a mean and an associated direction in feature space. We show that the sparse codes generated using this dictionary perform well in applications of recognition and reconstruction of human actions. Such a model also allows us to represent unseen actions accurately.

Extensions to non linear spaces: Features belonging to non linear spaces such as manifolds have become increasingly popular in the image processing and computer vision communities recently. An interesting extension to the proposed work could be to learn the proposed dictionary model on manifolds. Incorporating the non-linearity of the ambient space will lead to a model robust enough to work with these new features.

Compression of actions: With rising popularity of robots and intelligent surveillance systems, low bandwidth transmission for activities or events could prove to be extremely important. Using the proposed parametric form, extremely high compression ratios could be achieved since only the parameter(s) need to be transmitted as compared to several high dimensional features per action video.

5. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse

Representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *NIPS*, 2008.

[3] Wright, J. and Yang, A.Y. and Ganesh, A. and Sastry, S.S. and Yi Ma, "Robust face recognition via sparse representation," *IEEE Trans. on PAMI*, vol. 31, no. 2, pp. 210–227, 2001.

[4] T. Guha and R.K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on PAMI*, vol. 34, no. 8, pp. 1576 –1588, aug. 2012.

[5] Qiang Qiu, Zhuolin Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," in *ICCV*, nov. 2011, pp. 707 –714.

[6] Ahmed Elgammal and Chan-Su Lee, "Nonlinear manifold learning for dynamic shape and dynamic appearance," *Comput. Vis. Image Underst.*, vol. 106, no. 1, pp. 31–46, Apr. 2007.

[7] Zhaoshui He, Andrzej Cichocki, Yuanqing Li, Shengli Xie, and Saeid Sanei, "K-hyperline clustering learning for sparse component analysis," *Signal Process.*, vol. 89, no. 6, pp. 1011–1022, June 2009.

[8] J.J. Thiagarajan, K.N. Ramamurthy, and A. Spanias, "Optimality and stability of the K-hyperline clustering algorithm," *Pattern Recognition Letters*, vol. 32, no. 9, pp. 1299–1304, 2011.

[9] Sima Taheri, Pavan K. Turaga, and Rama Chellappa, "Towards view-invariant expression analysis using analytic shape manifolds," in *FG*, 2011, pp. 306–313.

[10] Troje N. F., "Decomposing biological motion : A framework for analysis and synthesis of human gait patterns," *Journal of Vision*, vol. 2, pp. 371–387, 2002.

[11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[12] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, "The function space of an activity," *IEEE CVPR*, pp. 959–968, 2006.

[13] Anuj Srivastava, Eric Klassen, Shantanu H. Joshi, and Ian H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1415–1428, 2011.

[14] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on PAMI*, vol. 29, no. 12, pp. 2247 –2253, dec. 2007.

[15] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *CVPR*, June 2009, pp. 1932 –1939.