Probabilistic subspace-based learning of shape dynamics modes for multi-view action recognition

S. Karthikeyan, Utkarsh Gaur, B.S. Manjunath
Department of Electrical and Computer Engineering,
University of California, Santa Barbara
karthikeyan@ece.ucsb.edu, utkarsh.gaur@yahoo.com, manj@ece.ucsb.edu

Scott Grafton
Department of Psychology
University of California, Santa Barbara
grafton@psych.ucsb.edu

Abstract

We propose a human action recognition algorithm by capturing a compact signature of shape dynamics from multi-view videos. First, we compute $\mathcal{R}$ transforms and its temporal velocity on action silhouettes from multiple views to generate a robust low level representation of shape. The spatio-temporal shape dynamics across all the views is then captured by fusion of eigen and multiset partial least squares modes. This provides us a lightweight signature which is classified using a probabilistic subspace similarity technique by learning inter-action and intra-action models. Quantitative and qualitative results of our algorithm are reported on MuHAVi a publicly available multi-camera multi-action dataset.

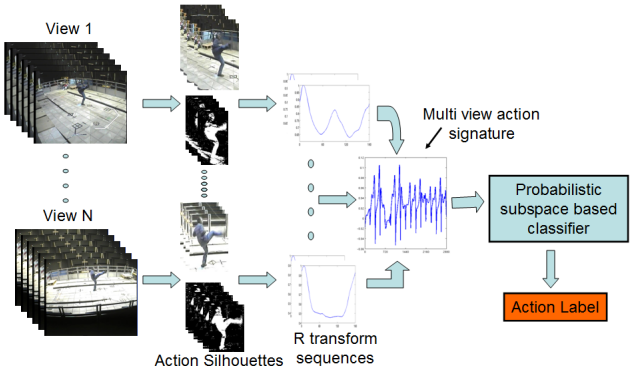

Figure 1. Flow diagram of our action recognition system

1. Introduction

Video cameras have become ubiquitous in all walks of life in the last decade. Several applications such as content-based video annotation and video summarization require recognition of actions occurring in videos. Action recognition also directly impacts surveillance and security. In this regard, researchers are focusing their attention on multi-view action recognition as fixed views are insufficient to discriminate large classes of actions. In addition, multi-view action recognition provides robustness to self and partial occlusions.

Several techniques have been proposed to tackle multi-view action recognition. Weinland et al. [1] proposed location and rotation invariant Fourier descriptors in cylindrical co-ordinates and compared two actions based on their 3D visual hull information. Yan et al. [2] proposed an arbitrary view action recognition system using 4D action feature model. Farhadi et al. [3] computed invariant features between two views for action discrimination. Similarly, Souvenir and Babbs [4] learned the viewpoint manifold of primitive actions to obtain a view-point invariant low dimensional representation of actions. Graphical model based methods [6, 5] are also popular for multi-view action recognition.

All these methods have limitations in terms of practical application. Most of the above mentioned techniques involve the construction of 3D visual hulls from multiple views. However, in realistic scenarios of uncalibrated camera networks without good multi-camera synchronization, the 3D visual hull construction is often unreliable. Additionally, these methods assume that the entire video sequence from all the cameras has to be transmitted to a central server. This practice incurs significant communication bandwidth consumption especially in multi-camera scenar-

ios. To overcome these problems, we propose a multi-view action recognition algorithm which works well in uncalibrated and unsynchronized networks, has low computational complexity and requires low communication bandwidth.

Shape and its deformation are important cues for human action recognition. Also, shape deformation implicitly captures motion information. Therefore we first compute frame level shape description of action silhouettes using $\mathcal{R}$ transform similar to the works of [7, 4]. In addition, shape deformation is captured using $\mathcal{R}$ transform temporal velocities.

From the frame level shape description we want to obtain a compact shape dynamics signature for the entire action sequence. Recently Ali et al. [8] proposed an eigenmode based representation for optical flow features which captures the primary variations of a single-view action sequence. We leverage this approach to build a compact multi-view action signature which incorporates both the intrinsic variations in individual views and co-occuring patterns across different views.

We capture the intrinsic variation of the shape dynamics in every view using the primary eigenmode. The dynamics across multiple views is represented by a novel descriptor based on multiset partial least squares(M-PLS) modes. These mode based features are computed both on the $\mathcal{R}$ transform and $\mathcal{R}$ transform temporal velocity. Subsequently, we obtain a compact signature for the entire multi-view video sequence. Next, as we typically observe that inter-action and intra-action signatures vary with different patterns, we build probabilistic models to understand these variations. Therefore, we adopt a probabilistic subspace similarity based classification technique which learns inter-action and intra-action subspace densities to predict the action label. The overall flow diagram of our approach is shown in Figure 1.

The following are the primary contributions of our work:

- Frame level shape deformation description using $\mathcal{R}$ transform and $\mathcal{R}$ transform temporal velocity.
- A novel compact multi-view signature to characterize shape dynamics across multiple views using eigen and M-PLS modes of $\mathcal{R}$ transform and $\mathcal{R}$ transform velocities.
- Learning inter-action and intra-action models and classifying a new action using a probabilistic similarity measure.

In addition, the $\mathcal{R}$ transform descriptor is 180 dimensional and transmitting this feature to a centralized server instead of the entire frame significantly reduces communication bandwidth.

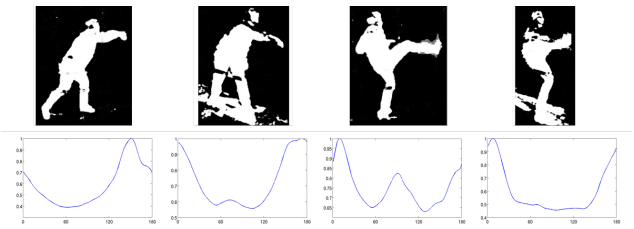

Figure 2. Example R Transforms of Punch and Kick actions for two camera views

2. Shape dynamics representation

2.1. R Transform

The $\mathcal{R}$ transform is a shape descriptor [7, 4] which converts a silhouette image into a compact 1D signal by using 2D radon transform. The 2D radon transform is the line integral of a line parameterized by (ρ, θ) in the polar coordinates. For an image frame $\mathbf{I}(x, y, t)$ at time t of a video clip $\mathbb{V}$ the radon transform, $r(\rho, \theta)$ is defined as

$$r(\rho, \theta, t) = \sum_x \sum_y \mathbf{I}(x, y, t)\delta(x\cos\theta + y\sin\theta - \rho) \quad (1)$$

where δ is the Dirac delta function.

The $\mathcal{R}$ transform is calculated as the sum of squared radon transform values for all lines corresponding to the same angle θ

$$\mathcal{R}(\theta, t) = \sum_\rho r^2(\rho, \theta, t) \quad (2)$$

The $\mathcal{R}$ transform is translation invariant and is robust to noisy silhouettes. However, to make it achieve some degree of scale invariance we use the normalized $\mathcal{R}$ transform in our experiments which is expressed as

$$\mathcal{R}'(\theta, t) = \frac{\mathcal{R}(\theta, t)}{\max_\theta(\mathcal{R}(\theta, t))} \quad (3)$$

We note that Figure 2 illustrates an example of $\mathcal{R}$ transforms of Punch and Kick actions for two different camera views.

2.2. R transform velocity

$\mathcal{R}$ transform provides us a salient representation of shape, but to characterize shape deformations from silhouettes we propose a novel feature using the velocity of the $\mathcal{R}$ transform. This is mathematically represented as

$$V(\theta, t) = \frac{d\mathcal{R}'(\theta, t)}{dt} \quad (4)$$

3. Mode based Shape dynamics signature

$\mathcal{R}$ transforms and its velocity are frame based feature descriptors. In this section, we obtain a signature for the entire multi-view activity sequence using these features. Mode based signatures for single view action recognition was first proposed by [8]. But, single view provides insufficient information to recognize a large class of actions. Hence, we propose an approach which develops a mode based action signature across multiple views. Our signature is obtained by fusing the eigen modes and M-PLS modes across different views.

3.1. Eigen mode

Principal component analysis (PCA) is a well known technique for determining an optimal basis for data reconstruction. Let $\{\mathbf{u}_k(\mathbf{x}, t_i)\}, i = 1, ..., M$ represent a sequence of features ($\mathcal{R}$ transform or $\mathcal{R}$ transform velocities) of a video sequence computed in each frame i in the kth view. A data matrix $\mathbf{U}_k(\mathbf{x})$ for view k is obtained as $[\mathbf{u}_k(\mathbf{x}, t_1)...\mathbf{u}_k(\mathbf{x}, t_M)]$. PCA extracts time independent orthonormal basis $\phi_k^i, i = 1, 2, ...M$ vectors also known as eigen modes along the directions of maximum data variation. The top eigen mode for the kth view, ϕ_k^m is vector ϕ_k that maximizes the following function

$$\begin{aligned} \phi_k^m &= \arg\max_{\phi_k} \ \phi_k^T \mathbf{C}_{\mathbf{U}_k\mathbf{U}_k} \phi_k \\ &\text{subject to } \phi_k^T \phi_k = 1 \end{aligned} \tag{5}$$

where $\mathbf{C}_{\mathbf{U}_k\mathbf{U}_k}$ is the covariance matrix of $\mathbf{U}_k(\mathbf{x})$. The effectiveness of eigen modes for action recognition is illustrated in [8].

3.2. Multiset Partial Least squares (M-PLS) mode

In our approach we want to obtain a multi-view signature of the video sequence. The eigen modes capture the primary shape dynamics in a single view. However, the correlation of the shape dynamics across multiple views is ignored. So we extract a novel set of shape dynamics modes which jointly maximize the covariance across all the views. Hence we use a multiset partial least squares method similar to [9]. Let the multi-view video sequence of N views be represented as $\mathbf{U}_k(\mathbf{x}), k = 1, 2, ..., N$. We want to compute projection vectors $\{\phi_k^m\}$ for every view k which maximizes the sum of covariances in the lower dimensional embedding. This can be mathematically written as

$$\begin{aligned} \{\phi_k^m\} &= \arg\max_{\{\phi_k\}} \sum_k \sum_{k' \neq k} \phi_k^T \mathbf{C}_{\mathbf{U}_k\mathbf{U}_{k'}} \phi_{k'} \\ &\text{subject to } \sum_k \phi_k^T \phi_k = 1 \end{aligned} \tag{6}$$

where $\mathbf{C}_{\mathbf{U}_k\mathbf{U}_{k'}}$ is the cross-covariance between $\mathbf{U}_k(\mathbf{x})$ and $\mathbf{U}_{k'}(\mathbf{x})$. Hence, $\{\phi_k^m\}$ in the form $[\ \phi_1^{mT}\ \phi_2^{mT}\ ...\ \phi_N^{mT}\ ]^T$ is obtained by computing the eigenmode of $\mathbf{C}$.

$$\mathbf{C} = \begin{bmatrix} 0 & \mathbf{C}_{\mathbf{U}_1\mathbf{U}_2} & ... & \mathbf{C}_{\mathbf{U}_1\mathbf{U}_N} \\ \mathbf{C}_{\mathbf{U}_2\mathbf{U}_1} & 0 & ... & \mathbf{C}_{\mathbf{U}_2\mathbf{U}_N} \\ . & . & . & . \\ \mathbf{C}_{\mathbf{U}_N\mathbf{U}_1} & \mathbf{C}_{\mathbf{U}_N\mathbf{U}_2} & ... & 0 \end{bmatrix} \tag{7}$$

3.3. Overall shape signature of actions

For a video sequence, we compute its shape dynamics signature by fusing(concatenating) the eigenmodes and M-PLS modes of the frame level $\mathcal{R}$ transform and $\mathcal{R}$ transform velocity descriptions, across all views. For a single view we have two eigenmodes and two M-PLS modes giving a 720 dimensional feature as $\mathcal{R}$ transform and its velocity are each 180 dimensional vectors. Therefore, the overall feature size for k views is a $720k$ dimensional vector.

4. Probabilistic subspace similarity based classifier

We observe that the signature described in the previous section has high intra-action similarity from Figure 3. Hence, we use a classifier which leverages this by modeling intra-action and inter-action variations. We also note that as number of views increases, the signature dimension becomes large. To overcome the curse of dimensionality, a subspace based density estimation technique is preferred. We also observe that in typical action recognition datasets, the number of training actions in every class is relatively low. Therefore, directly modeling every action class generatively is ineffective. But, inter-action and intra-action modeling creates a lot of training samples for both the classes and the probabilistic modeling becomes feasible.

We use the probabilistic subspace similarity learning proposed by Moghaddam et al. [10, 11] to learn intra-action and inter-action models. Given two action signatures, $\mathbf{s}_1$ and $\mathbf{s}_2$ we compute the difference between the signatures $\mathbf{v} = \mathbf{s}_1 - \mathbf{s}_2$ to train these models. This probabilistic subspace based density of $\mathbf{v} \in R^N$ divides the vector space R^N into two complementary subspaces. The target density is decomposed into two parts: density in the principal subspace $\mathbf{F}$ and its orthogonal complement space $\bar{\mathbf{F}}$. The density in the principal subspace is obtained using the first M principal components $\mathbf{y} = \{y_i\}_{i=1...M}$. The complete optimal high-dimensional density estimate for class Ω (inter or intra action) can be expressed as a product of two independent marginal gaussian densities

$$\begin{aligned} P(\mathbf{v}|\Omega) &= \left[\frac{\exp\left(-\frac{1}{2} \sum_{i=1}^{M} \frac{{y_i}^2}{\lambda_i} \right)}{(2\pi)^{\frac{M}{2}} \prod_{i=1}^{M} {\lambda_i}^{1/2}} \right] \left[\frac{\exp\left(-\frac{\epsilon^2(\mathbf{v})}{2\rho} \right)}{(2\pi\rho)^{(N-M)/2}} \right] \\ &= P_F(\mathbf{v}) \hat{P}_{\bar{F}}(\mathbf{v}|\rho) \end{aligned} \tag{8}$$

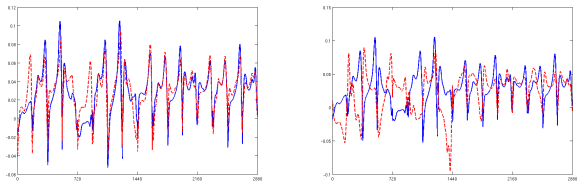

Figure 3. Left: signatures of Punch action corresponding to two different actors Right: signature of RunStop and Punch action by the same actor

Figure 4. The four views of WaveArms action used in our experiments

where $P_F(\mathbf{v})$ is the true marginal density in $\mathbf{F}$, and $\hat{P}_{\bar{F}}(\mathbf{v}|\rho)$ is the marginal density in the orthogonal complement space $\bar{\mathbf{F}}$. Here, $\epsilon^2(\mathbf{v})$ is the PCA residual and $\{\lambda_i\}$ are the top eigenvalues of the covariance matrix of $\mathbf{v}$. The optimal value of ρ is obtained by minimizing the divergence between the original probability density function and the approximation in (8). The optimal ρ is the average of the eigenvalues of $\bar{\mathbf{F}}$.

$$\rho = \frac{1}{N-M} \sum_{i=M+1}^{N} \lambda_i \tag{9}$$

Given a new test signature $\mathbf{s}_{te}$, every training signature is compared to the test signature by computing the probabilities of $\mathbf{v} = \mathbf{s}_{te} - \mathbf{s}_{tr}$ belonging to same action (Ω_s) and different action models (Ω_d) by using Bayes rule. Here, we assume the priors for both the classes are equal.

$$P(\Omega_s|\mathbf{v}) = \frac{P(\mathbf{v}|\Omega_s)P(\Omega_s)}{P(\mathbf{v}|\Omega_s)P(\Omega_s) + P(\mathbf{v}|\Omega_d)P(\Omega_d)} \tag{10}$$

The final class label is inferred by the training signature match which maximizes the score in equation 10.

5. Experiments

We present the evaluation of the proposed action recognition algorithm on the publicly available MuHAVi(Multicamera Human Action Video) dataset. We compare our method to [13] which uses bag of words(BoW) feature model on top of spatio-temporal interest point descriptors (STIP) [12] for multi-view action recognition.

5.1. Dataset

The MuHAVi dataset [14] consists of 17 action classes performed by 7 different actors. Among the current publicly available multi-view action datasets, MuHAVi contains the largest number of actions. The actions are WalkTurnBack, RunStop, Punch, Kick, ShotGunCollapse, PullHeavyObject, PickupThrowObject, WalkFall, LookInCar, CrawlOnKnees, WaveArms, DrawGraffiti, JumpOverFence, DrunkWalk, ClimbLadder, SmashObject and JumpOverGap. The actions were recorded by 8 different cameras, four sides and four corners of a rectangular platform. In our experiments we have used 4 of these camera views(V1, V3, V4 and V6) as shown in Figure 4. We note that the cameras have synchronization errors and calibration information is currently not available.

5.2. Results

In our analysis we used about 4 hours of action videos at 25 fps. We first preprocess the frames to obtain (noisy) silhouettes bounding the actor of interest. Then, we compute the $\mathcal{R}$ transform and obtain the entire action signature for all the actions by the 7 actors. These signatures are used to learn inter-action and intra-action models using the probabilistic learning technique described in Section 4.

We perform the testing in a "leave-one-actor-out" setting where, the probabilistic classifier was trained using all the videos except the actor corresponding to the test videos. In the training phase of this cross-validation technique, we can obtain at most 510 training samples to model intra-class variations. Hence, we use 510 samples for intra-class variations, and the inter-class variations are modeled using 510 randomly selected training samples. We note that the principal subspace is modeled in our classifier using 60 eigen-vectors. In this setting, we obtain a mean accuracy of 88.23% using our method and the confusion matrix across all the actions is shown in Figure 5.

We observe there is significant confusion between "Punch" and "SmashObject" actions. Both of these actions involve rapid hand movement justifying the misclassification. In addition, Figure 6 illustrates the recognition performance of our algorithm by changing the principal subspace dimension of our classifier. We observe that our results are quite consistent with change in size of the principal subspace with the best performance for 60 and 70 eigenvectors. We also note that our method outperforms [13] which uses Spatio-Temporal Interest Points (STIP) [12] for multi-view action recognition. This method constructs BoW of STIP features for every action. These BoWs are classified using Support Vector Machines and it obtains an average accuracy of 69.2% for "leave-one-actor-out" cross-validation. Figure 7 compares the performance of our proposed approach with [13] for every individual action.

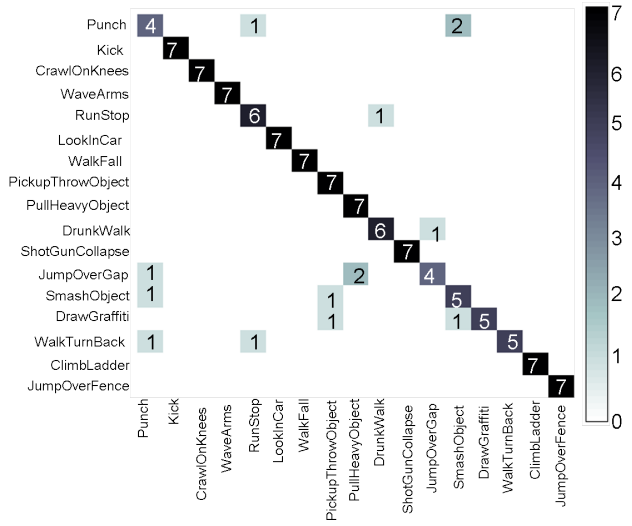

Figure 5. Confusion matrix for all the actions in MuHAVi dataset.

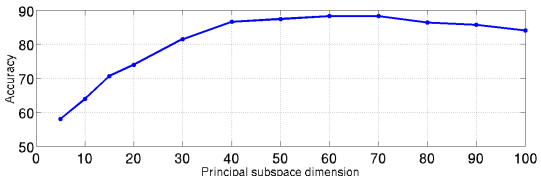

Figure 6. Action recognition accuracy of our method vs principal subspace dimension.

Our Method	57	100	100	100	86	100	100	100	100	86	100	57	71	71	71	100	100
STIP [13]	29	100	100	100	29	29	100	29	100	100	86	29	57	100	29	100	71
	Punch	Kick	CrawlOnKnees	WaveArms	RunStop	LookInCar	WalkFall	PickupThrowObject	PullHeavyObject	DrunkWalk	ShotGunCollapse	JumpOverGap	SmashObject	DrawGraffiti	WalkTurnBack	ClimbLadder	JumpOverFence

Figure 7. Comparison of performance of our method with [13] for every action.

6. Conclusions

In this work, we proposed a novel algorithm which encodes the static shape description and temporal shape dynamics to create a multi-view action signature. We modeled the inter-action and intra-action variations of this signature using a probabilistic subspace similarity-based classifier. Using this approach we performed impressively (88.2%) on MuHAVi, a complex multi-view action dataset which includes 17 actions.

7. Acknowledgments

We would like to acknowledge NSF award III-0808772, the MacArthur Foundation and Public Health Service grant NIMH 1 R01 MH070539-01. We also thank Dr. Sergio Velastin for providing access to the MuHAVi action recognition dataset.

References

[1] D. Weinland et. al, Free viewpoint action recognition using motion history volumes. *CVIU 2006*. 1

[2] P. Yan et. al, Learning 4d action feature models for arbitrary view action recognition. *CVPR 2008* 1

[3] A. Farhadi and M. Tabrizi, Learning to recognize activities from the wrong viewpoint. *ECCV 2008*. 1

[4] R. Souvenir and J. Babbs, Learning the viewpoint manifold of action recognition. *CVPR 2008* 1, 2

[5] D. Weinland et. al, Action recognition from arbitrary views using 3d exemplars. *CVPR 2007* 1

[6] P. Natarajan R. Nevatia, View and scale invariant action recognition using multiview shape-flow models. *CVPR 2008* 1

[7] Y. Wang et. al, Human activity recognition based on r transform. *CVPR 2007* 2

[8] S. Ali and M. Shah, Human action recognition in videos using kinematic features and multiple instance learning. *PAMI 2010* 2, 3

[9] A.A. Nielsen, Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *TIP 2002* 3

[10] B. Moghaddam and A. Pentland, Probabilistic visual learning for object representation. *PAMI 1997* 3

[11] B. Moghaddam, Principal manifolds and probabilistic subspaces for visual recognition. *PAMI 2002* 3

[12] I. Laptev, On space-time interest points. *IJCV 2005* 4

[13] C. Wu et. al, Multi-view activity recognition in smart homes with spatio-temporal features. *ICDSC 2010* 4, 5

[14] http://dipersec.king.ac.uk/MuHAVi-MAS/. 4