DISCRIMINATIVE SALIENCY-POSE-ATTENTION COVARIANCE FOR ACTION RECOGNITION

Jianhai Zhang *Zhiyong Feng* *Yong Su* *Meng Xing*

College of Intelligence and Computing, Tianjin University

ABSTRACT

Most covariance-based representations of actions are focused on the statistical features of poses by empirical averaging weighting. Note that these poses have a variety of saliency levels for different actions. Neglecting pose saliency could degrade the discriminative power of the covariance features, and further reduce the performance of action recognition. In this paper, we propose a novel saliency weighting covariance feature representation, Saliency-Pose-Attention Covariance(SPA-Cov), which reduces the negative effects from the ambiguous pose samples. Specifically, we utilize a discriminative approach to derive probability distribution of action categories for each pose, which is modeled by the uncertainty of information entropy to obtain the salient weighting. Experimental results show that our proposed method efficiently improves the discriminative power of the generated covariance. In some databases, the proposed SPA-Cov outperforms the state-of-the-art variant methods which are based on kernel matrix, Bayesian posterior features, temporal hierarchical features, etc.

Index Terms— Covariance, Pose, Attention, Saliency

1. INTRODUCTION

In recent decades, covariance matrix becomes very popular and is widely used in many computer vision tasks, such as speech processing[1], objective tracking[2], etc. This is due to the fact that it has two major advantages. Firstly, it provides more rich second-order statistical information than vanilla observations for various recognition tasks. Secondly, it is a Symmetric Positive Definite(SPD) matrix and lies in a Riemannian manifold, which has a well-developed theoretical property in mathematics.

For action time-series, covariance is used to model linear correlations between variables over temporal evolution[3]. Generally, empirical covariance is always used for feature representations. This means the poses in a sequence are equally weighted to generate covariance matrix. However, notice that the action poses have various saliency levels for different action categories[4]. For example, several actions could commonly include the standing pose, which satisfies the definition of ambiguous poses and is not very useful for some specific recognition tasks. Therefore, empirical covariance will degrade the performance of recognition tasks. This observation motivates us to pay more attention to some salient and discriminative poses, while neglecting some ambiguous poses[5]. In principle, enhancing discriminative information from salient poses and reducing useless, or even negative, information from ambiguous poses will make the generated covariances closer in a manifold space when they belong to the same action category, and farther when they belong to the different action categories.

In this paper, we propose a novel saliency weighting covariance feature representation, Saliency-Pose-Attention Covariance(SPA-Cov). Accordingly, a discriminative approach is proposed to utilize the uncertainty of information entropy to model the salient weighting of poses, which reduces the negative effects from the ambiguous pose samples and thus generates the discriminative rather than empirical covariance features. Our proposed SPA-Cov greatly improves the performance of the linear covariance matrix compared to empirical covariance. This is due to the fact that a discriminative guide knowledge of poses has been added into the generation process, unlike the other methods that just use self-information to generate feature representations. In the case of singularity, e.g. the number of statistical samples is less than feature dimensionality, we take a parametric conjugate prior of covariance to make it posterior, rather than simply appended a small scaled identity matrix[3]. Our method also can be extended to combine with the other methods. Just like the prior example above, making it more robust for feature representation.

The remainder of this paper is organized as follows: Section 2 introduces the related work. The detail of the SPA-Cov matrix is given in section 3. Experimental results are presented in section 4. Lastly, section 5 concludes the work in this paper.

2. RELATED WORK

In this paper, we propose an adapted covariance feature representation by discriminative saliency weighting of poses. Therefore, we will focus on two kinds of literature for action recognition, pose-aware and covariance-aware methods.

Pose-aware Methods. The effect of pose is crucially im-

portant for action recognition. The local features of poses can incur more widely applicable scenarios, e.g online recognition[4]. At first, action recognition engages in the recognition of still poses, recognizing actions by pose categories[6]. Subsequently, some methods of feature representation are proposed to improve the discriminative power of poses, e.g. space-time feature[7]. In [4], the authors conclude that the categories of the poses in actions have obvious discriminative saliency. This is because we just intend to pay more attention to the most useful poses in actions for classification[5]. To boost the power of saliency poses, some mining techniques of key poses are propose in [8]. In deep leaning domain, attention-based neural networks are proposed for action recognition[5]. Attention mechanism enables the learned network to automatically highlight the desired saliency information and learn a robust coding representation. For examples, [9] learns a similarity metric between sequences, which can align the concerned local information both temporally and spatially.

Covariance-aware Methods. Covariance matrix is first used as a region descriptor of image for classification[10]. Due to rich and robust statistic property, covariance is introduced to represent action time-series[3]. To improve the performance of covariance, some adapted methods are proposed for action recognition. For examples, in [11], the authors propose to map the covariances into the vectors in Euclidean space and conduct discriminative learning. [12] applies sparse coding and dictionary learning over SPD matrices, which further improve the expressive robustness. [13] suggests a third-order super-symmetric tensor representation instead of covariance. [14] proposes a tensor representation via kernel linearization, which will compactly capture higher-order relationships between body joints. In addition to the temporal correlation statistics, [15] imposes time-order into the covariance-based feature representations. To solve singularity and linear expressive power, [3] proposes non-linear kernel matrix to replace linear covariance, and [16][17] applies kernelized covariance, which maps pose samples into a infinite-dimensionality Hilbert space. [18] proposes the SPD dimension reduction on manifold to derive robust low-dimensional SPD matrices. In [19], the authors design a neural network to directly classify SPD matrices. In [20], the authors use locality projection on Riemannian manifold to find the most useful part in sequence.

3. PROPOSED METHOD

3.1. Preliminaries

Given an action sequence A $\in \mathbb{R}^{M\times T}$ is denoted by a pose set $\mathbf{D_A} = \{x_t\}_{t=1}^{T}$, the empirical covariance matrix involves the second-order linear statistical correlations among M feature variates:

$$\Sigma_{\mathbf{D_A}} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{F}(x_t - \mu) \tag{1}$$

where μ is a mean vector along the temporal axis and $\mathcal{F}(a) = a \otimes a \triangleq a \cdot a^{\mathbf{T}}$ denotes tensor product of vectors which will span to a symmetric second-order tensor.

From one point of view, the formula (1) is derived from the MLE of the parameters of Gaussian distribution, where the pose samples are assumed to follow a continuous Gaussian distribution $x_t \sim \mathcal{N}(\mu_{\mathbf{D_A}}, \Sigma_{\mathbf{D_A}})$. To perform MAP estimation, we have to specify conjugate priors about the parameters of this Gaussian distribution:

$$p(\mu_{\mathbf{D_A}}) = \mathcal{N}(\mu_{\mathbf{D_A}}|\mu_0, \Sigma_0) \tag{2}$$

$$\begin{aligned} p(\Sigma_{\mathbf{D_A}}) &= \mathcal{IW}(\Sigma_{\mathbf{D_A}}|S_0^{-1}, \gamma_0) &(3)\\ &\propto |\Sigma_0|^{-\frac{(M+\gamma_0+1)}{2}}\mathbf{exp}\big\{-\frac{1}{2}\mathbf{Tr}(S_0)\Sigma_0^{-1}\big\} &(4)\end{aligned}$$

where $\mathcal{IW}(\cdot)$ is a Inverse Wishart(IW) distribution, S_0^{-1} is a prior scatter matrix and $\gamma_0 > M-1$ is the degrees of freedom. We can obtain posterior by Bayesian formula, which also follows a IW distribution:

$$\begin{aligned} p(\Sigma_{\mathbf{D_A}}|\mathbf{D_A}, \mu_A) &\propto \mathbf{Prior}\times\mathbf{Likelihood} &(5)\\ &= \mathcal{IW}(\Sigma_{\mathbf{D_A}}|S_T^{-1}, \gamma_T) &(6)\\ S_T^{-1} &= S_0 + S_{\mu_{\mathbf{D_A}}} &(7)\\ \gamma_T &= \gamma_0 + T &(8)\end{aligned}$$

The posterior covariance can be obtained by the derivative of the log posterior function $\partial\mathbf{log}p(\Sigma_{\mathbf{D_A}}|\mathbf{D_A}, \mu_{\mathbf{D_A}})/\partial\Sigma_{\mathbf{D_A}}$:

$$\hat{\Sigma}_{\mathbf{D_A}} = \frac{S_T^{-1}}{M+\gamma_T+1} = \pi\Sigma_0 + (1-\pi)\Sigma_{\mathbf{D_A}} \tag{9}$$

where $\hat{\Sigma}_{\mathbf{D_A}}$ is a convex combination of prior and likelihood, and $\pi \in [0,1]$ controls the intensity of the prior knowledge. Note that the components of the posterior covariance, $\hat{\sigma}^{ij}_{\mathbf{D_A}} = \pi\sigma_0^{ij} + \frac{1-\pi}{T}\langle\bar{z}_i, \bar{z}_j\rangle$, still retain linear relationship, where $\bar{z}_i = z_i - \frac{1}{T}z_i \cdot \mathbf{1}_{T\times T}$ is the i-th row of the matrix $\mathbf{A}$ which is subtracted by its mean value.

3.2. Saliency-Pose-Attention Covariance Matrix

In order to derive a quantity describing saliency levels for pose samples in terms of action categories, we propose a non-parametric saliency weighting method. Ψ_{x_n} is a set including K most similar pose samples in terms of the pose x_n, which is obtained by K-NN algorithm:

$$\Psi_{x_n} = \mathbf{KNN}(x_n, \mathbf{D}^{\mathbf{train}}_{-x_n}) = \{(x_k, y_k)\}_{k=1}^{K} \tag{10}$$

where $\mathbf{D}^{\mathbf{train}}_{-x_n}$ is all the poses from the training set except the pose sample x_n, y_k is the action category identity, K is the number of the nearest neighbors. Therefore, we can count the probabilities $\mathcal{Z}_{x_n}$ for the pose x_n in terms of action categories as follows:

$$\mathcal{Z}^c_{x_n} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}(y_k = c), c = 1, \ldots, C \tag{11}$$

where $\mathbb{I}(\cdot)$ is an index function and the action category identity of each pose y_n follows a category distribution(multinoulli distribution) $y_n \sim \mathbf{Cat}(\mathcal{Z}_{x_n})$. Fortunately, the uncertainty of this distribution is able to reflect the saliency levels of the separate action categories for this pose x_n. The intuitive way to measure the uncertainty is to use information entropy towards a distribution. The maximal uncertainty corresponds to an uniform distribution, which means that this pose has the identical probability mass for allocating to the different action categories, thereby it conforms to a definition of the ambiguous pose. We thus want to suppress these ambiguous poses while augmenting the salient poses. For intuition, the entropy of the category distribution could be subsequently normalized as a saliency level due to boundaries of $\mathbb{H}(x_n) \in [0, -\mathbf{log}C]$:

$$\mathbb{H}(x_n) = -\sum_{c=1}^{C} \mathcal{Z}^c_{x_n} \mathbf{log} \mathcal{Z}^c_{x_n} \tag{12}$$

$$\mathcal{S}(x_n) = \frac{\mathbb{H}_{\mathbf{max}} - \mathbb{H}(x_n)}{\mathbb{H}_{\mathbf{max}}} \in [0, 1] \tag{13}$$

For the SPA covariance, the empirical average weighting is replaced by a normalized probability integral over an action sequence $\mathbf{D_A}$:

$$p(x_t|\mathbf{D_A}) = \frac{\mathcal{S}(x_t)}{\sum_{t=1}^{T} \mathcal{S}(x_t)} \tag{14}$$

Therefore, we are able to derive the SPA covariance $\Sigma^{\mathcal{S}}_{\mathbf{D_A}}$ by the statistical expectation of these second-order tensors:

$$\Sigma^{\mathcal{S}}_{\mathbf{D_A}} = \int_{x_t} p(x_t|\mathbf{D_A}) \mathcal{F}(x_t - \mu_{\mathcal{S}}) dx_t \tag{15}$$

where $\mu_{\mathcal{S}} = \int_{x_t} p(x_t|\mathbf{D_A}) \cdot x_t dx_t$ is the saliency weighted mean function.

In addition, we can also impose the regularized prior Σ_0 to against the case of singularity. In this paper, we take use of a very popular shrinkage estimate for MAP, where prior scatter matrix S_0 is set to $\mathbf{diag}(\Sigma^{\mathcal{S}}_{\mathbf{D_A}})$. The components of the posterior SPA covariance are given by:

$$\hat{\sigma}^{\mathcal{S},ij}_{\mathbf{D_A}} = \mathbb{I}(i = j) \cdot \sigma^{\mathcal{S},ij}_{\mathbf{D_A}} + \mathbb{I}(i \neq j) \cdot (1 - \pi)\sigma^{\mathcal{S},ij}_{\mathbf{D_A}} \tag{16}$$

where parametric π identically controls the prior intensity.

4. EXPERIMENTS

In this section, we evaluate the proposed SPA covariance on three databases.

MSR-Action3D(**S1**): 20 actions are performed by 10 subjects. The total number of the action sequences is 544.

MSR-DailyActivity3D(**S2**): 16 actions are performed by 10 subjects. The total number of the activity sequences is 320.

MSRC-Kinect12(**S3**): 12 gestures are performed by 30 subjects. 594 sequences and 719,359 frames are used. In total, there are 6,244 gesture instances.

4.1. Feature Representation of 3D skeletons

In the paper, we use the same configurations in [3][16][15], where only skeletal data is used. In **S1, S2**, each frame is represented by 120-dimensional differential velocity processed by [4]. In **S3**, each frame is used by 60-dimensional 3D-coordinate positions. The skeletal structures are uniformly normalized by preliminaries in [4]. The hyper-parameters are set by the cross-validations, where the prior penalty $\pi = 0.01$ in **S1-S3** and $K = 650, 800, 1000$ in **S1-S3**, respectively.

4.2. Classification Strategy

For a fair comparison, the implementation of the classification in our experiments is based on the available code provided by the authors in [3] and [16]. In the experiments, Multiple Kernel Learning(MKL) is applied to combine with different SPD matrix representations. The final classification is implemented by the SVM classifier. For the metric of SPD matrix, we use log-Euclidean Riemannian metric, which is easily calculated by $||\mathbf{log}(\mathbf{X}) - \mathbf{log}(\mathbf{Y})||_2$, where X and Y represent two SPD matrices. In addition, we respect the protocol that cross-subject test is used, where the odd-indexed subjects are used for training while the even-indexed subjects are used for testing.

4.3. Comparison Methods based on Covariance Matrix

In our experiments, six kinds of variants of the covariance-based feature representation are compared with our proposed SPA-Cov. The empirical covariance is denoted by Empirical-Cov for simplicity, and is calculated by (1). We compute the posterior covariance(Posterior-Cov) by (16). Infinite-Cov is measured by the metric of Bregman Divergences[17]. Temporal-Cov is implemented by the paper[15]. We adopt two kernels, RBF kernel and Polynomial kernel in [3], for comparison, which are named Kernel-RBF and Kernel-POL. Kernelized Covariance feature is denoted by Kernelized-Cov, which uses random fourier features in the paper[16].

The recognition accuracy of the compared methods are showed in **Table 1**. We can observe that the Empirical-cov has the worst performance compared to other covariance-based representations due to the linearity relationship and the

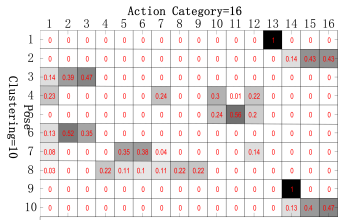
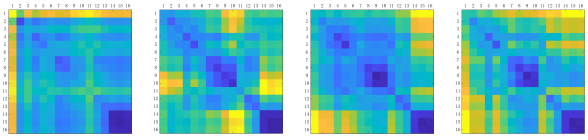


Fig. 1. Illustration of the significance of the saliency poses for recognition.



(a) (b) (c) (d)

Fig. 2. The average distances between 16 action categories with respect to (a) Empirical-Cov, (b) Kernel Matrix(RBF), (c) Kernelized-Cov, (d) proposed SPA-Cov in **S2** database.

sample scarcity. Infinite-Cov and Kernelized-Cov are generated by the similar process, but they take on the very different results. This is because that Kernelized-Cov applies the same MKL representation as in Kernel-RBF, whereas Infinite-Cov uses a Bregman Divergences to measure infinite-dimensional covariances and directly conduct classification by a SVM classifier. In performance, Posterier-Cov achieves the closely identical results with Kernel-based feature representations. This phenomenon means that the empirical covariance induced by a period of action sequence fails to express the real statistical features on a manifold space. We specially impose the prior knowledge to penalize covariance rather than append a scaled diagonal identity matrix to against singularity. Therefore, our Posterior-Cov greatly improves the recognition performance. From another point of view, our Posterior-Cov is equivalent to the Kernel methods when a specific form of multiplication of data matrices is given. In **Table 1**, Kernelized-Cov also achieves the same level of results. This is because that it essentially has the same discriminative information compared with Kernel-based representations.

In the proposed SPA-Cov, we allow for the significance of the saliency poses for recognition, which is illustrated in **Fig.1**. We use K-means algorithm to cluster pose samples in **S2**. Note that pose categories have various saliency levels for different action categories. For example, 8-th pose category has the large entropy about its category distribution, which means it has low-saliency for classification tasks, and thus needs the low-weighting for generating covariances. In contrast, 9-th is of zero-uncertainty so that it has high-weighting for generating covariances. As a result, the generated covariances will be closer in a manifold space when they belong to the same action category, and farther when they belong to the different action categories, as shown in **Fig.2**. In (d), the distances in off-diagonal positions are getting larger while the distances in diagonal are getting smaller compared to empirical covariance in (a). The SPA-Cov intuitively seems better than Kernel-Matrix and Kernelized-Cov. Moreover, the SPA-Cov representation still retains linearity relationship compared to Kernel-based Covariances, e.g. Kernel-RBF, Kernelized-Cov.

Methods/Databases	**S1**	**S2**	**S3**
Pose Set[21]	90.0	-/-	-/-
Moving Pose[4]	91.7	73.8	-/-
Empirical-Cov[10]	74.0	85.0	89.2
Infinite-Cov[17]	80.4	75.0	89.2
Temporal-Cov[15]	90.5	93.5	91.7
Kernel-RBF[3]	96.2	96.3	92.3
Kernel-POL[3]	**96.9**	96.9	90.5
Kernelized-Cov[16]	96.2	96.3	**95.0**
Proposed Posterior-Cov	96.2	96.3	89.5
Proposed SPA-Cov	96.2	**97.5**	90.5
Proposed Posterior-SPA-Cov	**96.9**	**97.5**	91.5

Table 1. Recognition accuracy (%) on **S1**, **S2**, **S3**. The sign of '-/-' means that no results are reported in the paper.

In **Table.1**, We can observe that the proposed SPA-Cov efficiently improves the discriminative power of the empirical covariance. Meanwhile, it also outperforms the state-off-the-art performance in **S2** and achieves the identical results in **S1**. However, in **S3**, the recognition accuracy of SPA-Cov is less than Kernelized-Cov. The reason might be that it only utilizes the linearity relationship or it fails to get a very effective saliency weighting. Therefore, we will carry on the study for some solutions and explore more expressive feature representations in the future.

5. CONCLUSIONS

The poses in actions have the various saliency levels for the different action categories. The empirical covariance generated by the averagely weighting poses will degrade the performance of recognition tasks. To solve this issue, we propose an approach selecting discriminative saliency poses to generate covariance, which is able to reduce the negative effects caused by the ambiguous poses according to the low-saliency levels. Experiments verify that the proposed SPA covariance efficiently improves the representation power of the generated covariance. Moreover, our approach is scalable that it can be extended to combine with the other methods, e.g. a prior knowledge is imposed to against singularity of sample scarcity. In the future work, we will explore more efficient methods to mine the saliency poses and extend it to much more cases, e.g. non-linear kernels, spatio-temporal representations, etc.

6. REFERENCES

[1] Ondrej Glembek, Jeff Ma, Pavel Matejka, Bing Zhang, Oldrich Plchot, Lukas Burget, and Spyros Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4032–4036.

[2] Xi Li, Weiming Hu, Zhongfei Zhang, Xiaoqin Zhang, Mingliang Zhu, and Jian Cheng, "Visual tracking via incremental log-euclidean riemannian subspace learning," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[3] Lei Wang, Jianjia Zhang, Luping Zhou, Chang Tang, and Wanqing Li, "Beyond covariance: Feature representation with nonlinear kernel matrices," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4570–4578.

[4] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2752–2759.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[6] Weilong Yang, Yang Wang, and Greg Mori, "Recognizing human actions from still images with latent poses," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2030–2037.

[7] Maxime Devanne, Hazem Wannous, Stefano Berretti, Pietro Pala, Mohamed Daoudi, and Alberto Del Bimbo, "Space-time pose representation for 3d human action recognition," in *International Conference on Image Analysis and Processing*. Springer, 2013, pp. 456–464.

[8] Chunyu Wang, Yizhou Wang, and Alan L Yuille, "Mining 3d key-pose-motifs for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2639–2647.

[9] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," *arXiv preprint arXiv:1708.02286*, 2017.

[10] Oncel Tuzel, Fatih Porikli, and Peter Meer, "Region covariance: A fast descriptor for detection and classification," in *European conference on computer vision*. Springer, 2006, pp. 589–600.

[11] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2496–2503.

[12] Mehrtash T Harandi, Conrad Sanderson, Richard Hartley, and Brian C Lovell, "Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach," in *Computer Vision–ECCV 2012*, pp. 216–229. Springer, 2012.

[13] Piotr Koniusz and Anoop Cherian, "Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5395–5403.

[14] Piotr Koniusz, Anoop Cherian, and Fatih Porikli, "Tensor representations via kernel linearization for action recognition from 3d skeletons," in *European Conference on Computer Vision*. Springer, 2016, pp. 37–53.

[15] Mohamed E Hussein, Marwan Torki, Mohammad Abdelaziz Gowayyed, and Motaz El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations.," in *IJCAI*, 2013, vol. 13, pp. 2466–2472.

[16] Jacopo Cavazza, Andrea Zunino, Marco San Biagio, and Vittorio Murino, "Kernelized covariance for action recognition," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 408–413.

[17] Mehrtash Harandi, Mathieu Salzmann, and Fatih Porikli, "Bregman divergences for infinite dimensional covariance matrices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1003–1010.

[18] Mehrtash T Harandi, Mathieu Salzmann, and Richard Hartley, "From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices," in *European conference on computer vision*. Springer, 2014, pp. 17–32.

[19] Zhiwu Huang and Luc J Van Gool, "A riemannian network for spd matrix learning," in *AAAI*, 2017, p. 3.

[20] Andres Sanin, Conrad Sanderson, Mehrtash T Harandi, and Brian C Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 103–110.

[21] Chunyu Wang, Yizhou Wang, and Alan L Yuille, "An approach to pose-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 915–922.