A STATISTIC MANIFOLD KERNEL WITH GRAPH EMBEDDING DISCRIMINANT ANALYSIS FOR ACTION AND EXPRESSION RECOGNITION

Shuanglu Dai and Hong Man

Department of Electrical and Computer Engineering,
Stevens Institute of Technology, Hoboken, NJ07030, U.S.A.

ABSTRACT

Graph embedding discriminant analysis is effective but computationally expensive for video-based recognition tasks. This paper proposes a statistic manifold kernel for visual modeling. Discriminant analysis can achieve effective computation with the proposed kernel for action and expression recognition. Firstly, symmetric positive definite (SPD) manifold is proposed to incorporate Gaussian mixture distribution of the video clips. Secondly, a projection kernel is constructed on the SPD manifold. Then an inter-class graph and an intra-class graph are introduced to measure the inter-class separability and intra-class compactness. The geometrical structure of the input data is thus exploited. A Marginal discriminant analysis(MDA) is finally performed on the kernel Hilbert space of the SPD Riemannian manifold. Recognition is achieved by the Nearest Neighbor (NN) method. Promising performances demonstrate the effectiveness of the proposed method for action and facial expression recognition.

Index Terms— Action recognition; Facial expression recognition; Marginal discriminant analysis; Statistic manifold kernel; Graph embedding

1. INTRODUCTION

Discriminant analysis on manifolds or kernel spaces has shown some effectiveness for image-based visual recognition and image set matching in recent works[1, 2]. Furthermore, graph embedding is introduced to explore the local similarities between visual features for better performances[3]. However, video clips with densely distributed images often requires high computational costs for discriminant analysis methods.

Many recent works have attempted to explore statistics for large-scale visual modeling and thus achieve effective computation. Statistics are able to represent one-shot clips temporally and object-centered clips spatially. Besides statistic representation, statistic-based metric learning methods such as Covariance discriminative learning[4] are introduced to learn discriminant subspaces. Furthermore, multiple statistics have been considered for visual modeling by the merit of metric learning. Huang et al. introduced a hybrid Euclidean-and-Riemannian distance function to incorporate multiple statistics on SPD Riemannian manifold[5]. With a similar distance function, Dai et al. further analyzed the canonical correlation of the kernel Hilbert spaces[6].

Compared with metric learning, discriminative distance learning is more efficient in learning and applying. Unfortunately, there are not many well-defined discriminant functions on multiple statistics. Typical discriminant learning methods such as Covariance discriminative learning(CDL)[4] and Manifold discriminant analysis(MDA)[2] are mostly learning discriminant functions with single statistics or in their kernel spaces. To improve the discriminant analysis for visual modeling, Harandi et al. explored a graph embedding method to incorporate the local similarities between visual features[3].

Inspired by the multiple statistic-based metric learning and graph embedding discriminant analysis, this paper proposes a statistic manifold kernel with graph embedding discriminant analysis for large-scale action and facial expression recognition. Firstly, the parameters of spatial Gaussian mixture distributions are generated on a symmetric positive definite (SPD) Riemannian manifold. By applying theory of Reproducing Kernel Hilbert spaces (RKHS), a projection manifold kernel is then constructed to measure the pairwise similarities. A graph embedding discriminant analysis is finally performed on the kernel Hilbert space. The inter-class graph and the intra-class graph introduced in this paper maintain the geometrical structure of the SPD manifold for discriminant analysis. The proposed method efficiently learns the discriminant space of the statistic manifold kernel. Promising results on action and facial expression recognition demonstrate the effectiveness of the proposed method.

2. GAUSSIAN MIXTURE STATISTICS AND ITS MANIFOLD KERNEL

Let$[X_1, ..., X_N]$ denotes N video clips, where X_i=$[x_1, ..., x_{n_i}]$$\in$ $R^{n_i \times d}$ indicates i-th clip, where n_i denotes the number of frames sampled from the i-th clip; d is the dimension of each frame in the clip; $1 \leq i \leq N$. The recognition task is to classify some input video clip to a discrete label $y \in \{1, .., L\}$.

Compute a n_i-dimension Gaussian mixture models(GMM) with H estimated mean vectors $\widehat{m}_h \in R^{1 \times n_i}, 1 \leq h \leq H$,

covariance matrices $\widehat{C}_i \in R^{n_i \times n_i}$ and multinomial priors w_h for the spatial distribution of video clip X_i, where $x_l \sim GMM_i(\widehat{m}_h, \widehat{C}_h, w_h)$ for $1 \leq l \leq n_i$.

Although the GMMs is representing video clips in both spatial and temporal, they are in different scales. A uniformed distance function is unable to measure the Gaussian mixture distributions in various scales. Inspired by [7] and [5], the symmetric positive definite (SPD) manifold is adopted to embed the parameter spaces of $GMM_i(\widehat{m}_h, \widehat{C}_h, w_h)$. The embedded SPD manifolds can thus be measured by Riemannian distances.

As it is proved in [6], the parameter space of a Single Gaussian Model(SGM) is able to be embedded into a $Sym^+_{n_i+1}$ space on the SPD Riemannian manifold. For the i-th clip X_i,

$$\Phi_i=\{SGM(\widehat{m},\widehat{C})| \begin{pmatrix} \widehat{C}+\widehat{m}\widehat{m}^T & \widehat{m} \\ \widehat{m}^T & 1 \end{pmatrix} \in R^{(n_i+1)\times(n_i+1)}, det(\Phi_i)>0\} \quad (1)$$

where Φ_i lies on a SPD Riemannian manifold. In order to measure the similarity $\forall(i,j)$ pairs, we introduce a projection manifold kernel $K_{i,j} = tr(log\Phi_i \cdot log\Phi_j)$, where $log(\cdot)$ denotes the SPD matrix logarithm. Let $k_i = tr(log\Phi_i \cdot log\Phi_j) \in R^{1\times N}, j \in [1,N]$ denote the i-th sample measuring the similarity from i-th clip to itself and other clips. Assume H Gaussians are estimated for all the video clips. According to the theory of Reproducing Kernel Hilbert Space (RKHS), the projection kernel of (i,j) pairs can be linearly combined as

$$K_{i,j}=\sum_{h=1}^{H}\sum_{h=1}^{H} w_i^h w_j^h tr(log\Phi_i^h \cdot log\Phi_j^h), (i,j)\in[1,N]\times[1,N] \quad (2)$$

where $\sum_{h=1}^{H} w_i^h=1, \sum_{h=1}^{H} w_j^h=1$ and $\sum_{h=1}^{H}\sum_{h=1}^{H} w_i^h w_j^h=1$. Eq.2 models N video clips into a projection SPD manifold kernel $K \in R^{N\times N}$, where each vector $k_i \in R^{1\times N}$ measures the similarities from the i-th manifold point to itself and other points.

3. GRAPH EMBEDDING DISCRIMINANT ANALYSIS ON SPD MANIFOLD KERNELS

Embed a graph $\mathcal{G}(V,W)$ where V denotes a collection of vertices or nodes and W denotes the collection of edges on the kernel space K. Let $V=\{v|v_i=k_i, i\in[1,N], v_i\in R^{1\times N}\}$ and $W\in R^{N\times N}$ to be a symmetric matrix where elements in W describe the linkages between vertices. The Laplacian matrix L of the graph $\mathcal{G}(V,W)$ is computed as $L = D - W$, where $D\in R^{N\times N}$ is a diagonal matrix computed as $D_{i,i}=\sum_{j\neq i} W_{i,j}, i \in [1,N], j \in [1,N]$.

Given kernel K with label space $Y=\{y_i, i\in[1,N]\}$ where y_i is the i-th class label for k_i. The intra-class similarity graph $W(\xi_{intra})$ and the inter-class similarity graph $W(\xi_{inter})$ are constructed by

$$W(\xi_{intra})=\begin{cases} 1, ||k_i-k_j||_2<\xi_{intra}, k_i,k_j\in K, y_i=y_j, \forall(i,j) \\ 0, otherwise \end{cases}$$

$$W(\xi_{inter})=\begin{cases} 1, ||k_i-k_j||_2<\xi_{inter}, k_i,k_j\in K, y_i\neq y_j, \forall(i,j) \\ 0, otherwise \end{cases} \quad (3)$$

where ξ_{intra} and ξ_{inter} indicate the radiuses of the intra-class and inter-class neighborhood, respectively. The L2 kernel distance measure $||k_i - k_j||_2$ in Eq.3 actually implies the Log-Euclidean distance $LED_{i,j} = ||log\Phi_i - log\Phi_j||_F$ for all the SPD manifold point pairs $(\Phi_i,\Phi_j),(i,j) \in [1,N]\times[1,N]$, where $||\cdot||_F$ denotes the Frobenius norm.

After graph embedding, a manifold kernel-based marginal discriminant analysis is derived. Assuming a linear mapping A on the kernel K is able to maximize the inter-class distance and minimize the intra-class distance, the two optimization problems are written as

$$max_A F_{inter}(A) = \frac{1}{2}\sum_{i,j}(Ak_i - Ak_j)^2 W(\xi_{inter}) \quad (4)$$

$$min_A F_{intra}(A) = \frac{1}{2}\sum_{i,j}(Ak_i - Ak_j)^2 W(\xi_{intra}) \quad (5)$$

where $F_{inter}(A) = AK^T(D(\xi_{inter}) - W(\xi_{inter}))KA^T = AK^T L(\xi_{inter})KA^T$; $L(\xi_{inter})$ is the inter-class Laplacian matrix. Similarly, $F_{intra}(A) = AK^T(D(\xi_{intra}) - W(\xi_{inter}))KA^T$ but with a constant term $AK^T(D(\xi_{intra})KA^T$. Since the kernel space for discriminant analysis is normalized in this paper, we force $AK^T D(\xi_{intra})KA^T = 1$. The intra-class minimization in Eq.(5) can be derived as

$$max_A AK^T W(\xi_{intra})KA^T, s.t. AK^T D(\xi_{intra})KA^T = 1 \quad (6)$$

Together with Eq.(4), the overall optimization problem can be considered as

$$\begin{aligned} &max_A AK^T(\gamma L(\xi_{inter}) + W(\xi_{intra}))KA^T \\ &s.t. AK^T D(\xi_{intra})KA^T = 1 \end{aligned} \quad (7)$$

where γ is a regularization parameter between the inter-class and intra-class distances. The Lagrangian function of A is

$$\begin{aligned} L(A) = &AK^T(\gamma L(\xi_{inter}) + W(\xi_{intra}))KA^T \\ &+\lambda(1 - AK^T D(\xi_{intra})KA^T) \end{aligned} \quad (8)$$

where λ is the Lagrangian multiplier. λ and A can be solved in form of an eigen-decomposition, where λ are the eigen-values and the optimal A obtained by the r-largest eigenvectors is in the form of Rayleigh quotient

$$A = \frac{K^T D(\xi_{intra})K}{K^T(\gamma L(\xi_{inter}) + W(\xi_{intra}))K} \quad (9)$$

After the optimal A being applied on kernel K, Nearest Neighbor (NN) is used for recognition in the discriminative space. The computation procedure of the graph embedding discriminant analysis on the statistic manifold kernel is summarized in Algorithm 1. For a conceptual understanding of algorithm 1, a space mapping structure of the proposed method for facial expression recognition is given in figure 1.

4. EXPERIMENTS

4.1. Recognition Tasks, Databases and Comparative Studies

In order to evaluate the robustness of the proposed method for different types of human-centered video recognition, two tasks are evaluated: action recognition and facial expression recognition.

The proposed method is compared with two types of methods for action recognition: 1) Statistic-based methods;

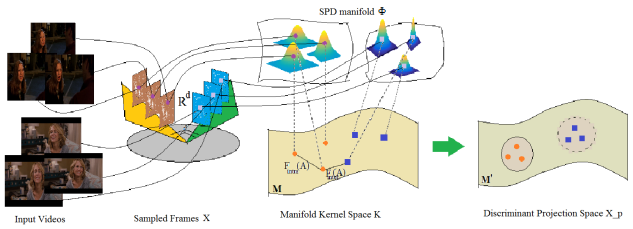


Fig. 1. Space mapping structure of the graph embedding discriminant analysis on the manifold kernel for facial expression recognition. Firstly, saliency visual parts are detected for every image, i.e. facial parts for expression and bodies for action, which lies in R^d linear subspaces. Secondly, the parameters of their GMDs are embedded on SPD manifold M. Distance functions $F_{intra}(A)$ and $F_{inter}(A)$ with embedded similarity graphs are then introduced for marginal discriminant analysis(MDA). The inter-class points are more compact and inter-class points are more separable than those before on a new manifold M'.

Algorithm 1 Graph embedding discriminant analysis on SPD manifold kernels

Training:

Input N sets of video frames $X=[X_1, ..., X_N]$ and labels $Y=[y_1, ..., y_N]$, regularization parameter γ, radiuses of the inter-class neighbor ξ_{inter} and intra-class neighbor ξ_{intra}

Compute N manifold points $[\Phi_1, ..., \Phi_N]$ by Eq.(1)

Compute kernel $K_{i,j}, \forall(i,j)\in[1,N]\times[1,N]$ by Eq.(2)

Compute intra-class graph $W(\xi_{intra}) \in R^{N\times N}$ and inter-class graph $W(\xi_{inter})\in R^{N\times N}$ by Eq.(3)

Compute $D(\xi_{intra})$, $D(\xi_{inter})$ and inter-class Laplacian matrix $L(\xi_{inter})=D(\xi_{inter})-W(\xi_{inter})$

return Projection matrix $A \in R^{N\times N}$ by Eq.(9) and Projected sample space $X_p = K^T A$

Testing:

Input N_{te} sets of video frames $X_{te} = [X_1, ..., X_{N_{te}}]$

Compute N_{te} manifold points $[\Phi_1, ..., \Phi_{N_{te}}]$ by Eq.(1)

Compute kernel $K^{te}_{i,j}$ with the N training manifold points, $\forall(i,j)\in[1,N]\times[1,N_{te}]$ by Eq.(2)

Compute projection $X^{te}_p = (K^{te})^T A$

Perform NN between X^{te}_p and X_p with label Y.

return Predicted label Y^{te}

2) Spatio-Temporal (ST) feature based methods. ST feature is a set of special features for action videos. The comparison with the latest ST-feature based methods will evaluate the competitiveness of the proposed method in the specific applications. The statistic-based methods are separated by different types of statistics: 1) Sample-based methods: Maximum Mean Discrepancy (MMD)[8]; 2) Subspace-based methods: Covariance Discriminant Learning (CDL)[4] and Manifold Discriminant Analysis (MDA)[2]; 3) Distribution-based methods: Gaussian Mixture Model (GMM)[9] and Single Gaussian Model (SGM)[10]; 4) Hybrid statistic methods: Statistic adaptive metric Learning (SAML) and Hybrid Euclidean-and-Riemannian metric Learning (HERML), where HERML and SAML are spatially extracting deep features for each frame. Some latest ST-feature based methods were proposed by Jhuang et al.[11], Wu et al.[12], Kovashka et al.[13], Liu et al.[14], Zhen et al.[15] and Simonyan et al.[16], where a convolutional neural network(CNN) is proposed by Simonyan et al. for action recognition. Five benchmark human action databases are used: Weizmann, UCF101, KTH, YouTube and HMDB51.

For facial expression recognition, the methods for comparison also fall into 2 categorises: 1) statistic-based methods; 2) facial feature-based methods. Statistic-based methods for facial expressions include: MDA, SAML and HERML. Latest facial feature-based methods include: Expressionlets (Explets) [17], PLSGrass[18]; HMM[19], CSPL[20], MSR[1], TMS[21] and Cov3D[7], where PLSGrass and Explets are proposed to use features extracted by CNN for each frame. Three benchmark facial expression databases are used: CK+, MMI and Acted Facial Expressions in the Wild(AFEW).

The average accuracy over 20 trials of classification are computed for performance evaluation. For the comparative methods, we adopt the reported default parameter settings. ξ_{intra}=20, ξ_{inter}=20 and γ=1 are adopted for algorithm 1 in the performance evaluation. Facial part detection and human body detection are performed before the kernel computation for facial expression and action recognition, respectively.

4.2. Performance Evaluations

4.2.1. Action Recognition

Table 1 shows the average accuracy of the proposed method compared with other statistic-based methods. From Table 1, we observe that the proposed method has highest accuracies among all methods on UCF101, HMDB51 and YouTube datasets. Table 2 shows the average accuracy of the proposed method compared with ST-feature based methods. As table 2 shows, the proposed method achieved the highest accuracy on KTH, UCF101, HMDB51, YouTube and the second highest accuracy on Weizmann. The higher accuracies achieved by the proposed method compared to the other ST-feature based methods show the advantage of the proposed kernel K for temporal feature representation in action recognition.

Table 1. Average recognition rate (%) compared with statistic-based methods on 3 action databases.

Statistic methods		UCF101	YouTube	HMDB51
Sample-based	MMD[8]	73.5	51.2	32.4
Subspace-based	CDL[4]	93.5	67.7	70.3
	MDA[10]	88.9	65.3	63.3
Distribution-based	GMM[9]	87.5	61.2	32.7
	SGM[10]	82.3	50.3	26.4
Hybrid	SAML[6]	95.5	70.8	71.3
	HERML[5]	92.5	74.6	68.1
proposed	DiscRie	**96.3**	**83.6**	**72.5**

4.2.2. Facial Expression Recognition

Table 3 shows the average accuracy of the proposed method compared with other statistic methods. From table 3, the proposed method achieved the highest accuracies among all methods on CK+, MMI and AFEW data sets. Table 4 shows the average accuracy of the proposed method compared with facial feature based methods. As table 4 shows, the proposed method also achieved the highest accuracy on the three data sets.

Table 2. Average recognition rate (%) compared with ST feature based methods on 5 action datasets.

ST-based	KTH	Weizman	UCF101	HMDB51	YouTube
Jhuang[11]	91.7	98.8	-	-	-
Wu[12]	94.5	-	91.3	-	-
Kovashka[13]	94.5	-	87.3	-	-
Liu[14]	95	-	-	48.4	82.3
Zhen[15]	94.5	-	80.7	31.7	80.7
Simonyan[16]	-	-	88.0	59.4	-
SAML[6]	95	91.25	95.5	71.3	70.83
proposed	**97.5**	**98.8**	**96.3**	**72.5**	**83.6**

Table 3. Average recognition rate (%) compared with statistic-based methods on 3 facial expression

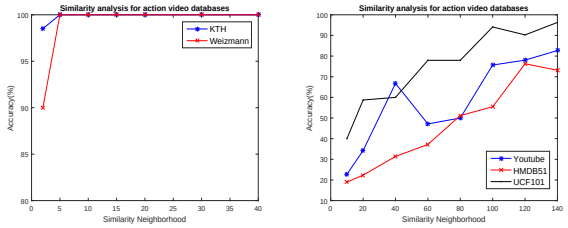
Statistic methods	CK+	MMI	AFEW
MDA[10]	73.5	51.2	22.4
SAML	81.25	70.8	30.3
HERML[5]	90.5	74.6	50.1
proposed	**93.25**	**83.6**	**53.5**

4.3. Performance Analysis

In the proposed method, the radiuses of inter-class and intra-class neighbor is changing according to different scenarios. This section further analyzes the performances with different levels of inter-class and intra-class similarities. Figure 2.a and 2.b show the accuracy of the proposed method with different levels of neighborhood similarities on 5 action benchmark video datasets. The accuracy in Figure 2.a show that the local similarities in KTH and Weizmann videos are quite small. The proposed methods has robust performance on such well-controlled action video data. In Figure 2.b, the 3 action video datasets were collected in the real-world scenarios and the accuracy become higher when similarities increase. It is evident that high local similarities exist in these video clips. Figure 3.a and 3.b show the accuracy of the proposed method with different levels of neighborhood similarities on 3 benchmark facial expression video data sets. The high accuracy in Figure 3.a show that the local similarities in CK+ are quite small and the classification is robust. In Figure 3.b, the 2 facial expression image data sets were collected in the real-world scenarios and movies. Again the accuracies become higher when similarities increase.

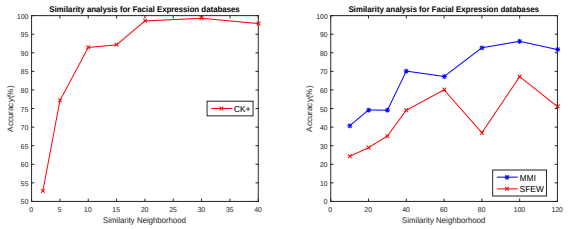
Table 4. Average recognition rate (%) compared with latest feature based methods on 3 facial expression datasets.

Feature-based	CK+	MMI	AFEW
HMM[19]	-	51.5	27.2
CSPL[20]	-	73.53	-
MSR[1]	91.4	-	-
TMS[21]	91.89	-	-
Cov3D[7]	92.3	-	22.04
Explets[17]	**94.19**	75.12	31.73
PLSGrass[18]	-	-	35.85
proposed	**93.25**	**83.6**	**53.5**



(a) KTH, Weizmann (b) UCF101, Youtube, HMDB51

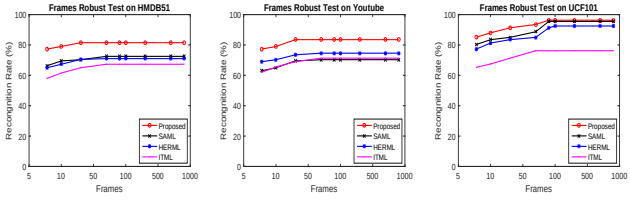
Fig. 2. Classification rates achieved by different similarity neighborhood on 5 benchmark action video datasets.



(a) CK+ (b) MMI and SFEW

Fig. 3. Classification rates achieved by different similarity neighborhood on 3 benchmark facial expression video datasets.

Other than parameter robustness, an evaluation of frame robustness is given by figure 4 in terms of average recognition rate w.r.t. the number of frames used for statistic computation. The proposed method is compared with 3 statistic-based metric learning methods: SAML, HERML and ITML on 3 real-world action video data sets. From figure 4, we observe that all the comparative metric learning methods reach their stable recognition accuracy with nearly the same speed. The proposed method is robust to the number of frames used per clip, but it is not able to reach stable recognition with less frames per clip.



(a) HMDB51 (b) Youtube (c) UCF101

Fig. 4. Evaluation of the frame robustness in different statistic-based methods on 3 real-world action video data sets.

5. CONCLUSIONS

This paper proposed a statistic manifold kernel with graph embedding discriminant analysis for human-centered video recognition. Statistics from GMMs are introduced to incorporate both spatial and temporal information of the video clips. The proposed method validated the graph embedding discriminant analysis for large-scale video recognition and outperformed main existing methods on action recognition and facial expression recognition. Future work will explore the similarity of the proposed kernel for different types of video recognition.

6. REFERENCES

[1] R. Ptucha, G. Tsagkatakis, and A. Savakis, "Manifold based sparse representation for robust expression recognition without neutral subtraction," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2136–2143.

[2] R. Wang and X. Chen, "Manifold discriminant analysis," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 429–436.

[3] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2705–2712.

[4] R. Wang, H. Guo, L. S. Davis, and Q. 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.

[5] Z. Huang, R. Wang, S. Shan, and X. Chen, "Hybrid euclidean-and-riemannian metric learning for image set classification," in *Computer Vision–ACCV 2014*. Springer, 2015, pp. 562–577.

[6] S. Dai and H. Man, "Statistical adaptive metric learning in visual action feature set recognition," *Image and Vision Computing*, 2016.

[7] A. Sanin, C. Sanderson, M. T. Harandi, and B. 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.

[8] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[9] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 581–588.

[10] M. Tistarelli and E. Grosso, "Identity management in face recognition systems," in *Biometrics and Identity Management*. Springer, 2008, pp. 67–81.

[11] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. Ieee, 2007, pp. 1–8.

[12] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 489–496.

[13] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2046–2053.

[14] L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: A genetic programming approach," *Cybernetics, IEEE Transactions on*, vol. 46, no. 1, pp. 158–170, 2016.

[15] X. Zhen, L. Shao, and X. Li, "Action recognition by spatio-temporal oriented energies," *Information Sciences*, vol. 281, pp. 295–309, 2014.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[17] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749–1756.

[18] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen, "Partial least squares regression on grassmannian manifold for emotion recognition," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 525–530.

[19] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3422–3429.

[20] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2562–2569.

[21] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1642–1649.