uc3m | Universidad Carlos III de Madrid

e-Archivo

This is a postprint version of the following published document:

Nazir, S., Yousaf, M.H y Velastin, S.A. (2017). Inter and Intra class correlation analysis (IIcCA) for human action recognition in realistic scenarios. In *8th International Conference of Pattern Recognition Systems (ICPRS 2017).*

DOI: https://doi.org/10.1049/cp.2017.0149

© 2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Inter and Intra Class Correlation Analysis ($II_C CA$) for Human Action Recognition in Realistic Scenarios

S.Nazir*, M.H.Yousaf*, S.A.Velastin+

*University of Engineering and Technology Taxila, Pakistan
+Universidad Carlos III de Madrid, Spain

Keywords: Human action recognition, inter and intra class variation, correlation analysis, UCF Sports

Abstract

Human action recognition in realistic scenarios is an important yet challenging task. In this paper we propose a new method, Inter and Intra class correlation analysis ($II_C CA$), to handle inter and intra class variations observed in realistic scenarios. Our contribution includes learning a class specific visual representation that efficiently represents a particular action class and has a high discriminative power with respect to other action classes. We use statistical measures to extract visual words that are highly intra correlated and less inter correlated. We evaluated and compared our approach with state-of-the-art work using a realistic benchmark human action recognition dataset.

1 Introduction

Human action recognition is a commonly studied area in computer vision. Its expansion took off in the early 1980s [1]. A wide-ranging literature exists about action recognition in a number of fields, including computer vision, signal processing, machine learning, pattern recognition etc. Human action recognition has been studied for more than a decade and its importance has grown since, due to its applications in human safety and security, including video surveillance, human computer interaction, robot learning etc. Many authors have made efforts to review and classify different approaches as well as cite different useful applications [2, 3, 4, 5, 6].

Action recognition is a challenging task due to a substantial amount of variation in video data. Realistic environments contain challenges like cluttered background, scale and view point variation, occlusion, variation in subject appearance etc. [7]. Recognition performance decreases in complex and realistic environments [8].

In addition to these challenges, variation within different classes impact recognition performance. For instance, two different individuals walking will display differences in terms of stride length and speed. In addition, similarities between two different actions classes (e.g. jogging can be considered as walking at a higher speed) can lead to confusion. As shown in Fig.1 the difference between both actions is very small. In particular, variation in actions performed by different subjects with different gender and at different speed and style needs to be handled [9]. Actions that seem so different and contain well-defined gestures according to us, can vary when performed in an uncontrolled environment. Thus, a major challenge is to deal with the large variations in action classes. These inter and intra class variations have been reported in many papers [10, 11]. In this paper, the ultimate goal is to propose a generic system with higher discriminative power to have a clear separation amongst these variations.

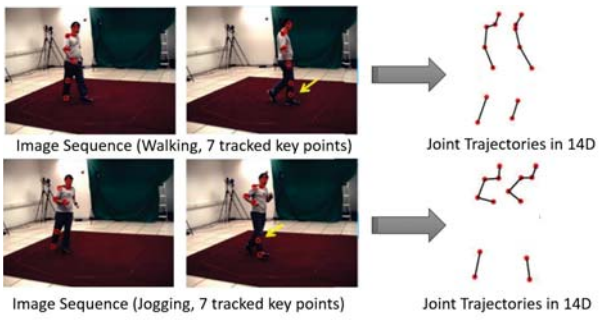

Figure 1. Walking and jogging action from the HumanEva videos dataset with their joint trajectories in 14D. (Courtesy [9]).

Wang et al. [12] proposed the actionlet ensemble method to handle the intra class variation in human actions. They captured human object interaction using LOP (Local Occupancy) features. Further, they represented each action as an actionlet: a conjunction of features for a subset of joints. They used a data mining approach to learn actionlets that are discriminative enough for each action class representation. They evaluated their approach on MSR-Action3D and MSRDaily Activity3D datasets. To handle inter and intra class variation Zunino [13] proposed a two-level classification approach. They recognize an action performed by a specific subject after identifying the subject first. They introduced a new evaluation strategy, known as personalization, to learn how actions are performed by a specific subject. Statistical measures were used to retrieve the subject specific role for handling inter and intra class variation challenge. They achieve reasonable performance on MSR-Action3D, MSRC-Kinect12 and HDM-05 benchmark datasets.

Here, we propose a new approach to analyze inter and intra class similarities between different action classes using correlation analysis. We aim to learn a model which can improve the

knowledge of our system by training it with class specific properties. We enrich our system with the information that is similar within a class and have a high discriminative power with respect to other action classes. During training, we select the adequate discriminative visual words representation for a particular action class and ignore the visual words with lower discriminative power using correlation analysis. As a result, our proposed approach provides significant results on a state-of-the-art realistic dataset for human action recognition.

2 II_CCA for Human Action Recognition

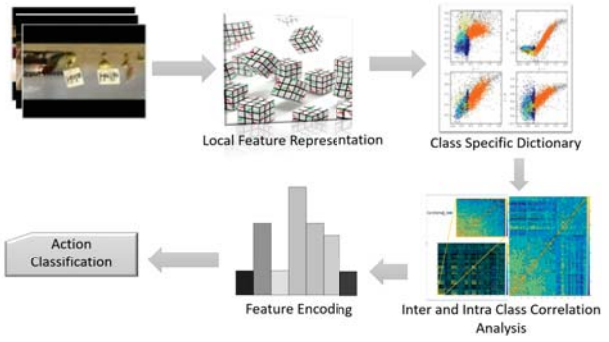

Figure 2. Proposed framework: II_CCA for human action recognition in realistic scenarios.

In this section, we discuss the proposed framework for realistic scenarios. Our approach follows a popular approach for human action recognition i.e. Bag of Words [14]. Our main focus is handling inter and intra class similarity challenges for human action recognition in a realistic environment. As shown in Fig.2 we represent each video using local feature representation, followed by the generation of class specific dictionaries for each action class. We use these class specific visual words representation for handling inter and intra class similarities using correlation analysis. After selecting highly intra correlated and less inter correlated visual words, we concatenate them into a visual word codebook. In the next step we encode video features using the visual words codebook and train a supervised classifier for action classification. The proposed approach is discussed in more detail in the following subsections.

2.1 Local feature representation

The first step in action recognition is to extract and represent features that are discriminative with respect to visual appearance and body movement of human body. To illustrate our approach, we use 3D Harris space time interest point detector as proposed by Laptev [15]. We use this detector to obtain interest points that are well localized in spatio-temporal domain and corresponds to meaningful events. These space time interest point corresponds to the non-constant motion in space time neighborhood. As proposed by Laptev [15], we have not adapted these interest point to scale and velocity so as to get sufficient interest points for video representation. These detected interest points are further described using a 3D SIFT descriptor [16]. 3D SIFT provides robustness to noise and orientation by encoding information in both space and time domain. It should be pointed out that the proposed method is independent of the feature representation chosen. Let $feat_i=\{f_1,f_2,f_3,....,f_p\}$ represent the total feature extracted for the i^{th} action class and p is the total number of features. These extracted features are grouped together as $FSet=\{feat_i\}^c{}_{i=1}$ where c is the total number of unique action classes.

2.2 Supervised class specific dictionary

The next step is to learn a dictionary that is discriminative enough to differentiate between different action class representations. Consider the feature representation of videos from each action class grouped together as $FSet=\{feat_i\}^c{}_{i=1}$ where c is the total number of unique action classes. We intend to learn a concatenated dictionary which consists of c class specific dictionaries $\phi_1, \phi_2, \phi_3, \ldots, \phi_c$.

Class specific dictionary learning has the advantage of class specific learning independently of other action classes [17]. It also have an advantage of parallel implementation as compared to classical dictionary learning technique [18]. Each class dictionary will have an *efficient* representation of its specific class and will be *less efficient* for other action classes.

2.3 Correlation analysis

2.3.1 Intra class correlation analysis

This section focuses on the correlation analysis of variation within each action class. Visual words from each action class ϕ_i are compared with themselves to obtain the relation between them. Let ϕ_i denote the visual words for the i^{th} action class. There are many ways to represent the correlation between different variables e.g. Pearson, Kendall and Spearman correlation. Pearson, Kendall and Spearman correlation coefficients are used to represent the association between two linear, ordinal and non-linear variables respectively. Because of the non-linear relationship, we measure the correlation between different visual words using the Spearman coefficient [19]. Spearman Correlation coefficient is computed as:

$$corr(\phi_i) = 1 - \frac{6d_w^2}{n(n-1)} \quad (1)$$

Where $d_w = r_g(\phi_i(w)) - rg(\phi_i(w))$ is the difference between two ranks of each visual word w and n is the total number of visual words in the i^{th} action class. The resultant matrix is denoted as $Corr(\phi_i)$for the i^{th} action class as shown in Fig.3(a). We consider this resultant matrix for mean correlation analysis. We used these statistics to characterize similarities and differences within a particular action class. Mean correlation is obtained by calculating the mean of all entries in $Corr(\phi_i)$ matrix and is represented as $MCorr(\phi_i)$. Mean correlation shows the degree of knowledge of each visual word within that class. The main diagonal elements of $Corr(\phi_i)$ shows the correlation of each visual word with itself, which is normally 1 as shown in Fig.3.

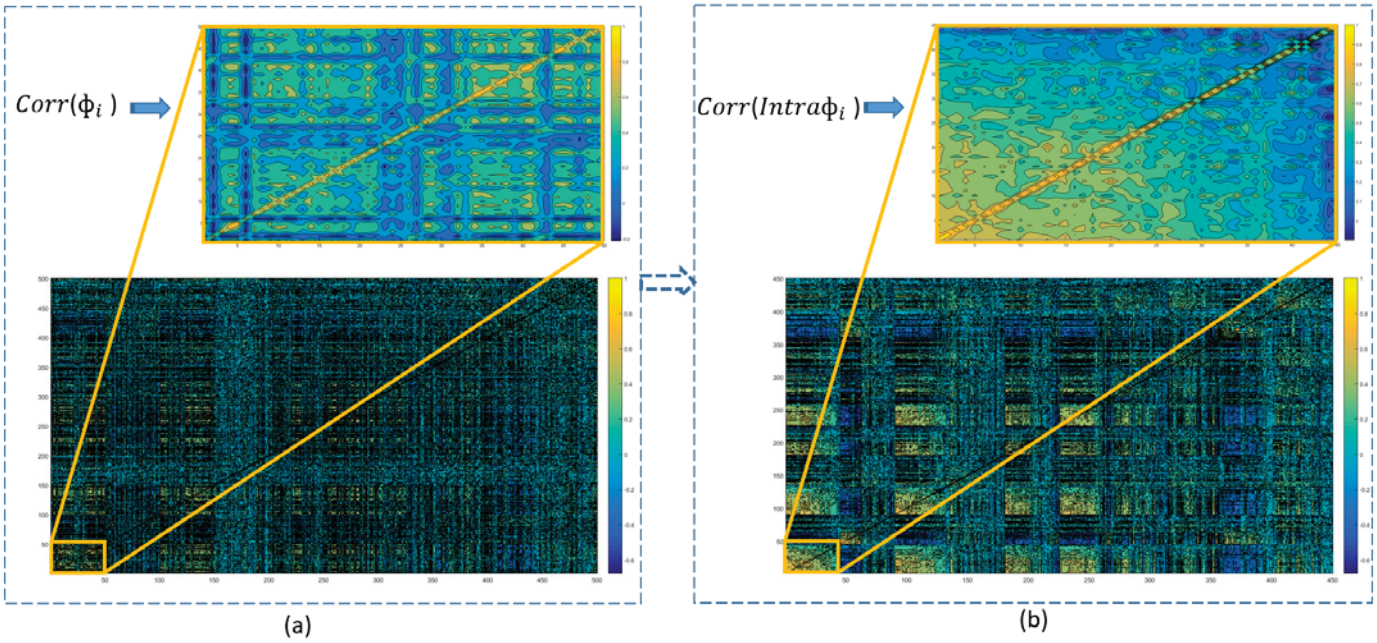

Figure 3. Correlation metrics for i^{th} action class (a) before intra class correlation analysis (b) after intra class correlation analysis.

The next step is the calculation of square root of mean correlation matrix. Square root of mean correlation shows the knowledge of each visual word among other visual words within its action class. Square root of mean correlation is denoted as $SMCorr(\phi_i)$ for i^{th} action class.

For maximizing intra class similarity, we have selected only those r visual words which are highly correlated with other visual words within its action class. In other words, we have only selected those visual words that efficiently represent its action class. Thus, the highly correlated visual words are denoted as $Intra\phi_i$ for the i^{th} action class. Correlation of these selected visual words is shown in Figure 3(b). High correlation is observed between the visual words representation of diving, kicking and riding horse action classes for UCF Sports dataset.

2.3.2 Inter class correlation analysis

In this section, we explain the measurement of variation between different action classes for inter class correlation analysis. We begin by computing the correlation between two different action classes. Let $Intra\phi_i$ and $Intra\phi_j$ represent the highly intra correlated visual words for the i^{th} and j^{th} action class respectively. We calculate the correlation between $Intra\phi_i$ and $Intra\phi_j$ using Spearman correlation coefficient computed as:

$$corr(Intra\phi_i, Intra\phi_j) = 1 - \frac{6d_w^2}{r(r-1)} \qquad (2)$$

Where $d_w = r_g(\phi_i(w)) - rg(\phi_j(w))$ is the difference between two ranks of each visual word w and r is the total number of highly intra correlated visual words in the i^{th} and j^{th} action class. The resultant matrix is denoted as Corr($Intra\phi_i$, $Intra\phi_j$). The resultant matrix consists of four quadrants, the upper right and lower left quadrant shows the correlation between $Intra\phi_i$ and $Intra\phi_j$. The upper left quadrant shows the correlation between $Intra\phi_i$ and the lower right quadrant shows the correlation between $Intra\phi_j$ as shown in Fig.4(a).

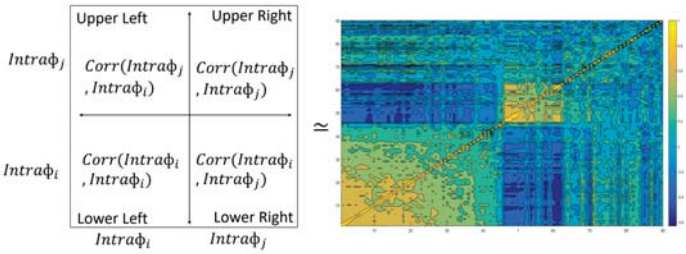

Figure 4. Correlation matrix for $Intra\phi_{ji}$ and $Intra\phi_j$.

As shown, using a graphical representation, in Fig.4(b) the upper right quadrant is the transpose of the lower left quadrant. We only choose the quadrant representing the correlation between $Intra\phi_i$ and $Intra\phi_j$ i.e. upper right quadrant or lower right quadrant for inter class correlation analysis. In the next step we calculate the mean correlation of these quadrants and denoted it as $MCorr(Intra\phi_i, Intra\phi_j)$. Followed by calculation of its square root, we represent it as $SMCorr(Intra\phi_i, Intra\phi_j)$.

Similarly to the concept discussed in the previous section, mean correlation matrix shows the degree of knowledge of each visual word of class i for action class j. The Square root of mean correlation shows the average knowledge of each visual word of i action class among visual words of j^{th} action class.

For minimizing the inter class similarity between different classes we have only selected those visual words of i^{th} action class which show less correlation with visual words of j^{th} action class. Here j varies from $1....c$ except i^{th} action class and c is the total number of unique action classes. In other words, we

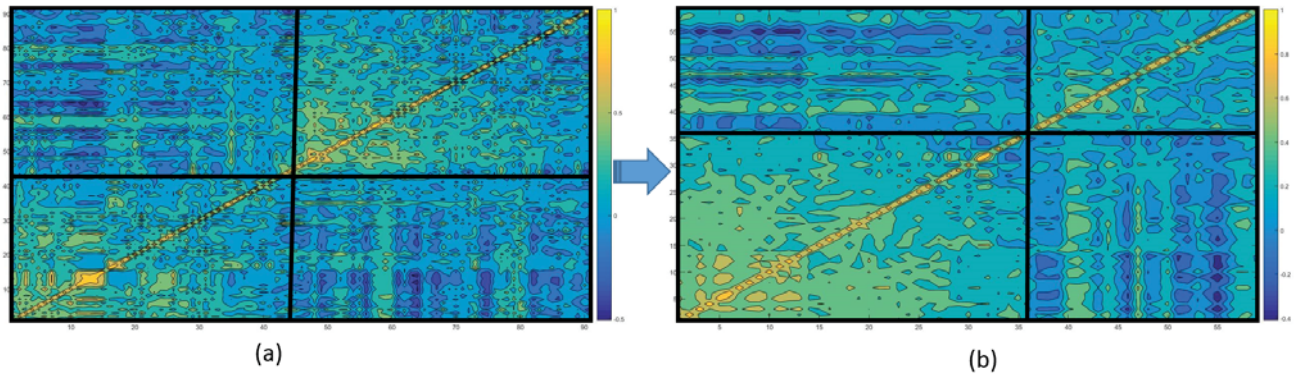

Figure 5. Correlation metrics the for i^{th} and j^{th} action classes (a) before inter class correlation analysis $Corr(Intra\phi_i, Intra\phi_j)$ (b) after inter class correlation analysis $Corr(Inter\phi_i, Inter\phi_j)$.

have only selected those visual words for the i^{th} action class that less efficiently represent the j^{th} action class. As a result, these selected visual words after inter class correlation analysis are represented as $Inter\phi_i$ for i[th] action class. Fig.5(b) shows the correlation of $Inter\phi_i$ and $Inter\phi_j$ class. The number of visual words for $Inter\phi_i$ are less than the number of visual words for $Intra\phi_i$ as shown in Fig.5(a).

2.3.3 Feature Encoding

After correlation analysis, we form a concatenated codebook $D=\{Inter\phi_1, Inter\phi_2, \ldots, Inter\phi_c\}$ where c is the total number of unique action classes. In this step, the main focus is encoding features for each video using the codebook D.

Let $F=\{f_1, f_2, f_3, \ldots, f_x\}$ represent the feature for each video. For each feature f_k the codebook word d_m can be viewed as a function of f and defined as

$$w(f) = \begin{cases} 1, & \arg\min_j ||f_k - d_j||_2 \\ 0, & otherwise \end{cases} \quad (3)$$

Where each feature vector votes for only its nearest codebook word. The occurrence of these votes are stored in histogram for each video.

2.4 Action Classification

Each video is represented as a histogram of highly intra class and less inter class correlated visual words. Further we train different classifiers using these video representation. We consider four different types of classification methods for training: Support Vector Machine (SVM), Nearest Neighbor Classifier (KNN), Decision Tree and Linear Discriminant analysis (LDA). For our experiments Linear discriminant analysis used empirical prior probabilities to determine class probabilities and KNN is trained by varying the number of nearest neighbors and $k=25$ provides the best result for as stated in Table.1. Decision tree considers $2^{c-1}-1$ combinations to predict the best split for class predictor where c is the total number of action classes. SVM used Gaussian kernel for learning which is defined as:

$$G(x_1, x_2) = exp(||x_1 - x_2||^2) \quad (4)$$

For all these classifiers we used the same cost measure which is defined as:

$$Cost(x, j) = \begin{cases} 1, & i \neq j \\ 0, & i == j \end{cases} \quad (5)$$

Our experimentation results shows that SVM performs better with respect to other three classifiers. SVM has also become a popular classifier for human action recognition. Our results show that Multiclass non-linear SVM trained using *c-1* binary support vector machine and the *'ordinal'* coding design scheme performs better, here c is the number of unique action classes.

3 Performance Evaluation

To test our approach, we performed number of experiments on publicly available dataset. All experiments were carried out on an Intel Core i7-6500U CPU with 2.50 Ghz, and the proposed algorithm was implemented in MATLAB 2015R(a).UCF Sports contains sports action videos captured in realistic environment. It contains 10 sports actions e.g. walking, diving, kicking, horseback riding etc. UCF Sports action videos have a large number of intra and inter class variation typical of many real life environments. We used leave one out cross validation method as proposed in [20].

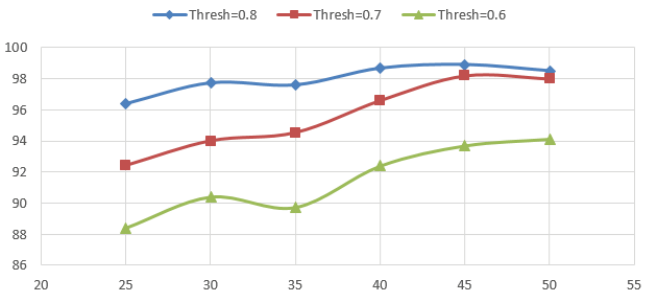

Figure 6. Parameter *(r and thresh)* evaluation for II_CCA.

For class specific dictionary construction we performed k-mean clustering using 50 visual words for each action class. In the next step, we performed different experiments to obtain

Figure 7. Confusion matrix for UCF Sports.

optimized *'r'* number of highly intra correlated visual words as show in Fig.6. Ignoring visual words representation for an action class by selecting smaller values of *r* can decrease performance. We further selected only those visual words that less efficiently represent other action class. We selected a threshold *'thresh*' by performing various experiments as shown in Fig.6. We have selected only those visual words that have less correlation with other action classes than this threshold.

Table 1. Classification method evaluation for II_CCA.

Classifier	**Accuracy**
Support Vector Machine	**98.90%**
K Nearest Neighbors	95.73%
Linear Discriminant Analysis	92.93%
Decision Tree	91.07%

Table 2. Comparison with state-of-the-art work for UCF Sports dataset.

Method	**Accuracy**
II_CCA (with SVM)	**98.90%**
CNN + Rank Pooling [21]	87.20%
Dense Trajectories + MBH [22]	88.00%
Independent sub space analysis [23]	86.50%

In our last experiment, we evaluated our approach using different classification methods for r=45 and thresh=0.8. As shown in Table.1, SVM proves to perform better with respect to other classifiers. Fig.7 shows the resultant confusion matrix for the UCF Sports dataset when evaluated using selected parameter as described above. As expected, significantly less inter and intra class similarity is observed between different classes for a realistic dataset (i.e. UCF Sports). Table.2 shows comparison with some other methods for human action recognition for UCF Sports dataset. Our approach achieves better performance as compared to other state-of-the-art methods.

4 Conclusion

In this paper, we have proposed a new approach to handle inter and intra class variation in realistic scenarios. We have shown that by computing the correlation between visual representations for each action class, we can handle inter and intra class variation challenge. First we learn class specific visual representation for each action class. Further we exploit these visual representations that have high intra class similarity and low inter class similarity. Finally, we demonstrate the potential of our proposed Inter and Intra class correlation analysis (II_CCA) approach for action recognition by evaluating its performance on a realistic human action recognition dataset. Future work will be to strengthen the robustness of our approach to other challenges like occlusion and view invariance for human action recognition in realistic scenarios.

Acknowledgements

S.A. Velastin has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nº 600371, the Ministerio de Economía, Industria y Competitividad (COFUND2013-51509) the Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander.

References

[1] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 20–27, IEEE, 2012.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[4] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[5] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[6] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

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

[8] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[9] D. Gong and G. Medioni, "Dynamic manifold warping for view invariant action recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 571–578, IEEE, 2011.

[10] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[11] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 588–595, 2014.

[12] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1290–1297, IEEE, 2012.

[13] A. Zunino, J. Cavazza, and V. Murino, "Revisiting human action recognition: Personalization vs. generalization," *arXiv preprint arXiv:1605.00392*, 2016.

[14] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[15] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[16] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 357–360, ACM, 2007.

[17] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.

[18] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognition*, vol. 45, no. 11, pp. 3902–3911, 2012.

[19] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficient on the same sets of data," *Quaestiones Geographicae*, vol. 2, no. 30, 2011.

[20] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[21] B. Fernando and S. Gould, "Learning end-to-end video classification with rank-pooling," in *Proc. of the International Conference on Machine Learning (ICML)*, 2016.

[22] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[23] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3361–3368, IEEE, 2011.