A Non-negative Low Rank and Sparse Model for Action Recognition

Biyun Sheng[1], Wankou Yang[1], Baochang Zhang[2], and Changyin Sun[1]

[1] School of Automation, Southeast University, Nanjing 210096, China
[2] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
hisby@126.com, wankou.yang@yahoo.com, bczhang@buaa.edu.cn, cysun@seu.edu.cn

Abstract. In this paper, we present a new method for video action recognition. The main contributions are two-fold. First, we propose local coordinates contained descriptors (LCCD) instead of appearance-only descriptors. We encode global geometric correspondence by combining descriptors with spatio-temporal locations, which is different from previous methods such as spatio-temporal pyramid matching (STPM). Spatio-temporal location is taken as part of the coding step by utilizing LCCD. Second, a novel non-negative low rank and sparse coding model is developed to encode descriptors for action recognition. Motivated by low rank matrix recovery and completion, local descriptors in a spatio-temporal neighborhood are similar and should be approximately low rank. The objective function is obtained by seeking non-negative low rank and sparse coefficients for local descriptors. The learned coefficients can capture location information and the structure of descriptors, hence improve the discriminability of representations. Experiments validate that our method achieves the state-of-the-art results on two benchmark datasets.

Keywords: local coordinates, non-negative low rank, sparse coding, action recognition.

1 Introduction

In recent years, action recognition in videos has been a very active research area due to its wide applications such as in surveillance, human-computer interface, sports video analysis, and content based video retrieval [1]. State-of-the-art performances have been achieved by the Bag of Visual Words (BOVW) method, which includes extraction of local descriptors (e.g., HOG or HOF) and construction of representations.

In the framework of BOVW, the collection of unordered words ignores the interest points' location information. Aiming at the loss of location information,Choi et al. extend the spatial pyramid method for video retrieval and propose Spatio-Temporal Pyramid Matching (STPM) [6]. The concatenation of histograms leads to huge vector representation. The finer the region is portioned, the longer the final representation is. Yuan et al. introduce a new global

S. Li et al. (Eds.): CCPR 2014, Part II, CCIS 484, pp. 266–275, 2014.

feature called 3D R transform, which captures the distribution of interest points [1]. The global feature and the BOVW representation are then combined by a context-aware feature fusion method. The method improves the accuracy while it brings computational complexity.

Restrictive cardinality constraint on vector quantization in BOVW leads to relatively high reconstruction error. Sparse coding technique has attracted much attention to reduce the reconstruction error. However, it has a drawback that sparse codes cannot vary smoothly on the data manifold. The dependence of local descriptors is ignored which results in different codes for similar descriptors. Gao et al. propose Laplacian sparse coding to exploit the dependence among the local descriptors [2]. This algorithm preserves the consistence of sparse representation for similar local descriptors while the large number of descriptors leads to computational infeasibility as well as impracticality in real-world applications.

In fact, local spatially and temporally descriptors close in a video should have similar sparse codes ideally. The low rank representation can easily solve the non-consistency problem [4]. Promising results have been shown by low rank and sparse matrix recovery in many applications [3],[8]. However, limited work has applied the low rank sparse coding framework to solve action recognition problem.

Usually sparse coding technique or its variant is followed by max pooling to get the final representation. The sign of coding coefficients is not constrained traditionally. Negative coefficients appear in order to satisfy the objective function, while large numbers of zero coefficients are inevitable. Since non-zero components typically provide useful information, the max pooling process will bring the loss in terms of those negative components, and further degrade the classification performance [5]. Besides, it is meaningful to reduce the information loss by non-negative constraint on coding coefficients during the encoding process.

In this paper, we propose new descriptors called local coordinates contained descriptors (LCCD) and calculate corresponding coefficients by non-negative low rank sparse coding method. Fig.1 shows the flowchart of our framework. We encode coordinates of spatio-temporal interest points (STIPs) as well as the corresponding descriptors so that the representations themselves contain location information. For the encoding model, we add low rank regularizer and non-negative constraint into the traditional sparse coding objective function. The low rankness enforces similar descriptors to have similar sparse codes, which considers the local geometrical structure of the data. The non-negative constraint lowers information loss for representations. Therefore, the learned representations are remarkably more discriminative.

The remainder of this paper is organized as follows. Section 2 introduces the new descriptors with appearance feature and location information, called LCCD. Section 3 presents the non-negative low rank and sparse coding method. Section 4 experimentally tests our method on two human action datasets. Section 5 concludes the paper.

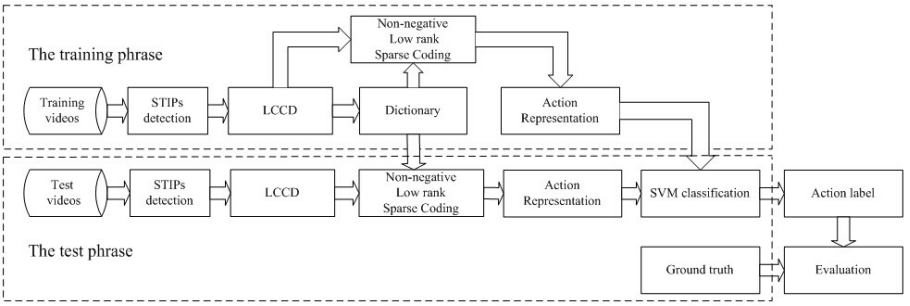

Fig. 1. Flowchart of the proposed action recognition framework

2 Local Coordinates Contained Descriptors (LCCD)

For a video, a set of interest points are detected and traditional descriptors are obtained based on every interest point. These descriptors own the local texture or motion information, however, ignores the location information of the interest points. Instead of utilizing more complicated descriptors such as 3D R transform [1] or settling the problem during pool stage by STPM [6], we propose a more intuitive and easier method to describe video descriptors, namely local coordinates contained descriptors (LCCD).

We first perform STIPs detection by the Harris operator. A multi-scale approach is used. The HOG/HOF feature is adopted to describe the cuboid extracted at each interest point [12]. LCCD of a video are denoted as:

$$\begin{cases} X = [X_1, X_2, \ldots, X_i, \ldots, X_N] \\ X_i = [\varphi_i; \alpha x_i; \alpha y_i; \beta t_i] \end{cases}, \quad 1 \le i \le N \tag{1}$$

where α and β are parameters with functions of coordinate normalization, location weight regulation and dimensional transformation,(x_i, y_i, t_i) is the coordinate of the i^{th} interest point,φ_i is the HOG/HOF feature, and N is the total number of interest points detected in the video.

In contrast to the original appearance-only descriptors, the proposed ones contain location information which is beneficial to capture geometric structure of the data. Compared with STPM, there is no need for dividing the video artificially to define the pooling regions. Appearance-only descriptors and their coordinates are simultaneously encoded so that the learned coefficients have more discriminative power.

3 Non-negative Low Rank Sparse Coding

In this section, we first introduce the non-negative low rank sparse model and then give the optimization process.

3.1 Non-negative Low Rank Sparse Model

In the sparse model, the input signal is well approximated by a sparse linear combination of the given overcomplete bases in dictionary. Such sparse representations are usually derived by linear programming as an l_1-norm minimization problem. But the l_1 based regularization is sensitive to outliers. Therefore, we use $l_{2,1}$-norm instead in this paper.

Suppose $X = [X_1, X_2, \ldots, \ldots, X_N] \in \mathbb{R}^{d \times N}$ be LCCD for a video, in which d, N respectively denote the dimension and number of descriptors. The proposed non-negative low rank sparse model is

$$\min_U \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|U\|_* + \lambda_2 \|U\|_{2,1} \qquad s.t.\ U \geq 0 \tag{2}$$

where $\|.\|_F, \|.\|_*, \|.\|_{2,1}$ respectively denotes the Frobenius-norm, the nuclear norm, and the $l_{2,1}$-norm of a matrix. $U = [U_1, U_2, \ldots, U_N]$ is the coefficient matrix with each U_i being the representation of X_i. The nuclear norm, a convex approximation to the rank function, is the sum of the singular values of a matrix. $\lambda_i (i = 1, 2)$ are parameters to trade off low rankness and sparsity.

From the proposed model in (2), we can find the fact that the model degenerates to the sparse coding model if we set the parameter $\lambda_1 = 0$. The nuclear norm here is used to enforce the codes of similar descriptors in neighborhood to be approximately similar. Fig.2 shows the comparison between standard sparse coding (SC) and our low rank sparse coding (LRSC). Different from SC, similar bases are selected to guarantee the consistency of similar descriptors in LRSC.

(a) SC (b) LRSC

Fig. 2. Comparison between SC and LRSC. X_1 and X_2 are two similar inputs to be encoded.

Without non-negative constraint, the coefficients learned by low rank sparse model can be negative. Zero (or small positive) coefficients indicate the corresponding bases in the dictionary have no (or very small) influence. However, since zero (or positive value) is always larger than negative values, max pooling strategy will choose zero (or positive value) instead of negative values [5]. It not only leads to worse performance for data representation, but also lacks physical interpretation for many visual data. Therefore, the non-negative constraint on the coefficients is meaningful and necessary.

3.2 Optimization Process

Inexact Augmented Lagrange multipliers (IALM) have been applied to solve the low rank problem [8]. We first introduce two auxiliary variable V and W to make regularizations of the objective function in (2) separable. The problem (2) can be transformed as follows:

$$\min_{U,V,W} \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|V\|_* + \lambda_2 \|W\|_{2,1} \qquad s.t.\ U = V, U = W, W \geq 0 \quad (3)$$

The augmented Lagrangian function of problem (3) is

$$\begin{aligned} L(U,V,W,Y_1,Y_2,u) &= \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|V\|_* + \lambda_2 \|W\|_{2,1} + < Y_1, U - V > \\ &+ < Y_2, U - W > + \frac{u}{2}(\|U - V\|_F^2 + \|U - W\|_F^2) \\ &= \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|V\|_* + \lambda_2 \|W\|_{2,1} \\ &+ h(U,V,W,Y_1,Y_2,u) - \frac{1}{2u}(\|Y_1\|_F^2 + \|Y_2\|_F^2) \end{aligned} \quad (4)$$

where

$$\begin{cases} h(U,V,W,Y_1,Y_2,u) = \dfrac{u}{2}(\|U - V + \dfrac{1}{u}Y_1\|^2 + \|U - W + \dfrac{1}{u}Y_2\|^2); \\ < A, B >= tr(A^T B) \end{cases} \quad (5)$$

The dictionary B in (3) is calculated by k-means. By the method of IALM, the objective function achieves convergence by a sequence of closed form update steps. The variable U,V or W is updated with other variables fixed. The updating schemes are as follows.

$$\begin{aligned} U_{k+1} &= \operatorname*{argmin}_U \frac{1}{2} \|X - BU\|_F^2 + < Y_{1,k}, U - V_k > \\ &+ < Y_{2,k}, U - W_k > + \frac{u_k}{2}(\|U - V_k\|_F^2 + \|U - W_k\|_F^2) \\ &= (B^T B + 2u_k I)^{-1}(B^T X - Y_{1,k} - Y_{2,k} + u_k V_k + u_k W_k) \end{aligned} \quad (6)$$

$$\begin{aligned} V_{k+1} &= \operatorname*{argmin}_V \frac{\lambda_1}{u_k}\|V\|_* + \frac{1}{2}\|V - (U_k + \frac{1}{u_k}Y_{1,k})\|_F^2 \\ &= \Theta_{\frac{\lambda_1}{u_k}}(U_k + \frac{1}{u_k}Y_{1,k}) \end{aligned} \quad (7)$$

$$\begin{aligned} W_{k+1} &= \operatorname*{argmin}_{W \geq 0} \frac{\lambda_2}{u_k}\|W\|_{2,1} + \frac{1}{2}\|(W - (U_k + \frac{1}{u_k}Y_{2,k}))\|_F^2 \\ &= max(\Omega_{\frac{\lambda_2}{u_k}}(U_k + \frac{1}{u_k}Y_{2,k}), 0) \end{aligned} \quad (8)$$

where Θ and Ω are respectively singular value soft-thresholding operator and $l_{2,1}$ minimization operator. In detail, the form of analytic solutions for Θ and Ω are as follows:

$$\Theta_\lambda(A) = U_A S_\lambda(\Sigma_A) V_A^T \tag{9}$$

In (9),$A = U_A \Sigma_A V_A^T$ is the SVD of A and $S_\lambda(A_{ij}) = sign(A_{ij}) max(0, |A_{ij}| - \lambda)$ is soft-thresholding operator.

Let $A = [a_1, a_2, \cdots, a_i, \cdots]$ be a given matrix, then the i^{th} column of $\Omega_\lambda(A)$ is $\frac{max(0, \|a_i\| - \lambda)}{\|a_i\|} a_i$.

Algorithm1. Non-Negative Low Rank Sparse Coding via IALM

```
Input:Data X, Dictionary B, and Parameters λ1 and λ2;
Output:U,V,W;
const
  ρ = 1.1;u = 0.1;maxiter = 10e30;ε = 10e-3;
var
  iter: 0..maxiter;
begin
  iter := 0;
  repeat
    fix V,W and update variable U according to (6);
    fix W,U and update variable V according to (7);
    fix U,V and update variable W according to (8);
    Y1,iter+1 := Y1,iter+u(Uiter-Viter);
    Y2,iter+1 := Y2,iter+u(Uiter-Witer);
    u=ρu;  iter := iter + 1;
   until ‖U − V‖∞ < ε and ‖U − W‖∞ < ε;or iter = maxiter;
end.
```

4 Experiments

We test our approach on two benchmark datasets: the KTH actions dataset [13], and the UCF Sports dataset [14].

4.1 Experiments on the KTH Dataset

The KTH action dataset contains six types of human actions (boxing, hand waving, hand clapping, walking, jogging and running), performed repeatedly by 25 subjects in four different scenarios: outdoors, outdoors with camera zoom, outdoors with different clothes and indoors. Twenty-four actors' videos are used as the training sets and the remaining one person's videos as the testing set. The results are the average of 25 times runs. We empirically set the size of the dictionary to 250 for the dataset. For the non-negative low rank sparse model, we set the tradeoff parameters $\lambda_1 = 1, \lambda_2 = 0.1$.

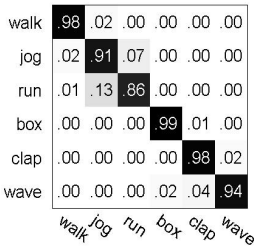

Fig. 3. Confusion matrix on KTH

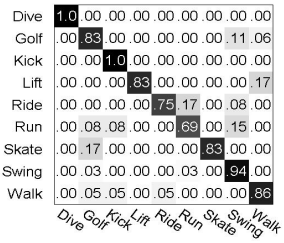

Fig. 4. Confusion matrix on UCF

Fig.3 shows the confusion matrix across all scenarios. The figure demonstrates the effectiveness of our proposed approach. For example, the accuracies of some actions such as "walking", "boxing" and "handclapping" can reach above 97%. The"running" action is easily misclassified as "jogging" because of the high similarity between the two actions. Table 1 lists the average accuracy of action recognition by other researchers.

Table 1. Comparison with previous work on the KTH dataset

Approach	Year	Accuracy(%)
Brendel et al.[17]	2010	94.22
Le et al.[7]	2011	93.90
Zhang et al.[16]	2012	95.5
Wang et al.[15]	2013	94.2
Ours		**94.32**

Compared with the listed results in recent research, our method achieves 94.32%, which is comparable to the state-of-the-art result. However, with the original appearance-only descriptors the recognition rate is 93.32%. The experimental result illustrates that the proposed descriptors improve the performance of our framework. Traditional descriptors don't utilize the location information of STIPs (or settle the problem during pooling stage), which leads to a set of unordered representations (or lengthy representations). Our descriptors contain location of the interest point and appearance characteristics of cuboid around the point.

In order to validate the effectiveness of non-negative constraint, low rank and $l_{2,1}$-norm regularizer, we change one of the above three terms with others fixed. The results of comparison are shown in Table 2. When we calculate absolute values of coefficient matrix instead of non-negative constraint, the accuracy is only 91.82%. It shows that taking absolute values artificially is not reasonable in our model and may drop the accuracy significantly. The performance of our

Table 2. Effectiveness of non-negative low rank sparse model

Method	Accuracy(%)
Without non-negative constraint	91.82
Without low rank regularizer	93.32
l_1-norm instead of $l_{2,1}$-norm	93.82
Ours	**94.32**

model without low rank regularizer is 93.32%. If we change $l_{2,1}$-norm to l_1-norm, the accuracy is 93.82% . In combination of the three terms, the accuracy of our method is 94.32%. The experimental results demonstrate that the representations obtained by our proposed method are more discriminative.

4.2 Experiments on the UCF Sports Dataset

The UCF Sports dataset consists of 150 videos with 9 action classes taken from real broadcasts (e.g., diving, golf swinging, kicking), with different viewpoints and scene backgrounds. The dataset is tested in a leave-one-out manner, cycling each example in as a test video one at a time. We empirically set the size of the dictionary to 800 for the dataset. For the non-negative low rank sparse model, we set the tradeoff parameters $\lambda_1 = 1, \lambda_2 = 0.1$.

Fig.4 shows the confusion matrix of our approach on the UCF dataset. The recognition rate for some actions is high up to 100% such as "Dive" and "Kick". Experimental results by previous methods are listed in Table 3.

Table 3. Comparison with previous work on the UCF dataset

Approach	Year	Accuracy(%)
Kovashka et al.[18]	2010	87.27
Le et al.[7]	2011	86.5
Yuan et al.[1]	2012	87.33
Wang et al.[15]	2013	88.0
Ours		**88.0**

When we use the traditional descriptors and pool the coefficients by STPM, the accuracy is about 80% which is much lower than our method. The result shows that the proposed LCCD is especially fit for the UCF dataset. We do the same experiments as in section 4.1 to validate our proposed model, and Table 4 illustrates the performances of different combinations. Non-negative constraint here is vital and effects the final result largely.

Table 4. Effectiveness of non-negative low rank sparse model

Method	Accuracy(%)
Without non-negative constraint	82.0
Without low rank regularizer	86.67
l_1-norm instead of $l_{2,1}$-norm	88.0
Ours	**88.0**

5 Conclusion

In this paper, we have presented a novel method to learn representations of human actions. In order to describe the "where" property of STIPs, we encode descriptors with location information. Besides, we adopt non-negative low rank sparse coding technique. The learned coefficients have the property of spatio-temporal consistency and finally boost the accuracy. Extensive experiments on two datasets have demonstrated the effectiveness of our proposed approach.

Acknowledgments. TThis work is supported by National Natural Science Foundation (NNSF) of China under Grant . 61375001, 61473086, partly supported by the open fund of Key Laboratory of Measurement and partly supported by Control of Complex Systems of Engineering, Ministry of Education (No. MCCSE2013B01), and the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology), (No. 30920130122006).

References

1. Yuan, C.F., Li, X., Hu, W.M., Ling, H.B., Maybank, S.: 3D R Transform on Spatio-Temporal Interest Points for Action Recognition. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, Portland (2013)
2. Gao, S.H., Tsang, I.W.H., Chia, L.T., Zhao, P.L.: Local features are not lonely - laplacian sparse coding for imageclassification. In: 23th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–2. IEEE Press, San Francisco (2010)
3. Zhuang, L.S., Gao, H.Y., Lin, Z.C., Ma, Y., Zhang, X.: Non-Negative Low Rank and Sparse Graph for Semi-Supervised Learning. In: 25th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2328–2331. IEEE Press, Providence (2012)
4. Zhang, T.Z., Ghanem, B., Liu, S., Xu, C.S., Zhang, X., Yu, N.H., Ahuja, N.: Low-Rank Sparse Coding for Image Classification. In: 14th IEEE International Conference on Computer Vision, pp. 281–286. IEEE Press, Sydney (2013)
5. Zhang, C.J., Liu, J., Tian, Q., Xu, C.S.: Image Classification by Non-Negative Sparse Coding, Low-Rank and Sparse Decomposition. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1673–1678. IEEE Press, Colorado (2011)

6. Choi, J., Wang, Z.Y., Lee, S.C.: Spatio-temporal pyramid matching for sports videos. In: Proceeding MIR 2008 Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 291–297. IEEE Press, New York (2008)
7. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3361–3368. IEEE Press, Providence (2011)
8. Zhang, Y.M.Z., Jiang, Z.L., Davis, L.S.: Learning Structured Low-rank Representations for Image Classification. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–3. IEEE Press, Portland (2013)
9. Jiang, Z.L., Ghanem, B., Liu, S., Ahuja, N.: Low-rank sparse learning for robust visual tracking. In: 12th European Conference on Computer Vision, pp. 470–474. IEEE Press, Florence (2012)
10. Zhang, Z.D., Matsushita, Y., Ma, Y.: Camera Calibration with Lens Distortion from Low-rank Textures. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2321–2328. IEEE Press, Providence (2011)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 19th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, New York (2006)
12. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning Realistic Human Actions from Movies. In: 21th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, Alaska (2008)
13. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: Proceedings of International Conference on Pattern Recognition, 17th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–4. IEEE Press, Washington (2004)
14. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 21st IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, Alaska (2008)
15. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Dense Trajectories and Motion Boundary Descriptors for Action Recognition. IJCV 103, 60–79 (2013)
16. Zhang, Y.M., Liu, X.M., Chang, M.C., Ge, W.N., Chen, T.: Spatio-Temporal Phrases for Activity Recognition. In: 12th European Conference on Computer Vision, pp. 707–721. IEEE Press, San Francisco (2012)
17. Brendel, W., Todorovic, S.: Activities as Time Series of Human Postures. In: 11th European Conference on Computer Vision, pp. 9–13. IEEE Press, Greece (2010)
18. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: 23rd IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–4. IEEE Press, San Francisco (2010)