Cholesky Decomposition-Based Metric Learning for Video-Based Human Action Recognition

SI CHEN1, YUANYUAN SHEN2, YAN YAN3, DAHAN WANG1, AND SHUNZHI ZHU1

1School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China
2Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
3School of Informatics, Xiamen University, Xiamen 361005, China

Corresponding author: Shunzhi Zhu (ssz@xmut.edu.cn)

This work was supported in part by the Natural Science Foundation of Fujian Province of China under Grant 2018J01576 and Grant 2017J01127, in part by the National Key Research and Development Program of China under Grant 2017YFB1302400, and in part by the National Natural Science Foundation of China under Grant 61672442, Grant 61773325 and Grant 61571379.

ABSTRACT Video-based human action recognition can understand human actions and behaviours in the video sequences, and has wide applications for health care, human-machine interaction and so on. Metric learning, which learns a similarity metric, plays an important role in human action recognition. However, learning a full-rank matrix is usually inefficient and easily leads to overfitting. In order to overcome the above issues, a common way is to impose the low-rank constraint on the learned matrix. This paper proposes a novel Cholesky decomposition based metric learning (CDML) method for effective video-based human action recognition. Firstly, the improved dense trajectories technique and the vector of locally aggregated descriptor (VLAD) are respectively used for feature detection and feature encoding. Then, considering the high dimensionality of VLAD features, we propose to learn a similarity matrix by taking advantage of Cholesky decomposition, which decomposes the matrix into the product between a lower triangular matrix and its symmetric matrix. Different from the traditional low-rank metric learning methods that explicitly adopt the low-rank constraint to learn the matrix, the proposed algorithm achieves such a constraint by controlling the rank of the lower triangular matrix, thus leading to high computational efficiency. Experimental results on the public video dataset show that the proposed method achieves the superior performance compared with several state-of-the-art methods.

INDEX TERMS Human action recognition, metric learning, Cholesky decomposition.

I. INTRODUCTION

Video-based human action recognition aims to recognize and understand the actions and behaviours in the video sequences. It has attracted increasing attention in computer vision, mainly because of its wide variety of applications, such as health care [1] and human-machine interaction [2]. For example, in the health care domain, the human actions are automatically recognized in real time and thus the corresponding medical rescue can be provided timely based on video-based human action recognition. In the report of literature [29], some actions, such as running, jumping, climbing and falling, are viewed as potentially dangerous for seniors and children, and action recognition based on video can greatly help to reduce the occurrence of accidents.

After intensive researches during recent years, the video-based human action recognition technology has made great progresses and developments. However, due to the dynamic appearance changes of the human [33] caused by the internal and external factors, for example, illumination changes, different attitudes, fast actions of the cameras, and the dramatic changes in actions, video-based human action recognition in complex environments still faces huge difficulties and challenges.

During the past few decades, the significant progress has been made and many methods have been developed on video-based action recognition [3]–[6]. These methods usually consider video-based action recognition as a supervised learning process, where the relationship between different actions is modelled via the training set, and then the learned model is used to recognize actions in the test set. For example, Subhransu et al. [4] first extracted a set of action-related
features, and then the action is classified by using Support Vector Machine (SVM). Maji et al. [5] used a trained body detector for human detection, and then employed a non-linear SVM for classification. Yao et al. [6] extracted the features by dense sampling, and then utilized random forest to distinguish different actions. The above methods are based on manually-designed features, such as Histogram of Oriented Gradient (HOG) features [7] and Scale-Invariant Feature Transform (SIFT) features [8].

Currently, several methods [9], [10], [28], [30] rely on deep learning to perform video-based action recognition. For example, Gkioxari et al. [9] proposed to use R*NN for context-based action recognition. Peng and Schmid [10] developed a multi-region two-stream R-CNN method for action detection. Chen et al. [28] proposed a spatiotemporal heterogeneous two-stream network for video action recognition, which employs two different network structures for spatial and temporal information, respectively. Tang et al. [30] proposed a Semantics Preserving Teacher-Student (SPTS) networks architecture, which is applied on action segmentation task. Although the excellent performance achieved by the deep learning based methods, these methods often require the large-scale training data, and thus the training complexity is relatively high [11]. In this paper, we focus on the small sample size problem (i.e., the training samples are limited), where the manually-designed features are exploited.

In general, the features extracted from the video sequences are in high dimensions. However, the high-dimensional features often contain significant redundant information and noises, which will cause the performance decrease of the subsequent classification. Therefore, metric learning (which usually learns a distance/similarity metric) can not only play the role of dimensionality reduction, but also improve the recognition rate by matching the similarity between features for video-based action recognition. Specifically, similarity metric learning or distance metric learning uses a training sample set to obtain a metric model that effectively reflects the similarity (or distance) between the training sample and the test sample, and thus matches the similar action to the same class and distinguishes the dissimilar actions to be the different classes.

The larger the distance between the training sample and the test sample becomes, the more dissimilar the two samples are. For example, Hu et al. [31] developed a sharable and individual multi-view metric learning approach for visual recognition, which jointly learns an optimal combination of multiple distance metrics on multi-view representations. Tran and Sorokin [32] applied the Large Margin nearest neighbor (LMNN) algorithm to behavior recognition, and used the metric learning algorithm to obtain a Mahalanobis distance (learning matrix in Mahalanobis distance) to improve the performance of human activity recognition.

The above metric learning methods [31], [32] learn a full-rank matrix, such that the similarities between the same-class samples become higher, while those between the different-class samples keep far apart with the learned matrix. Nevertheless, learning a full-rank matrix is often inefficient and is subject to overfitting. Hence, an effective way to overcome these problems is to impose the regularization (such as the sparse or low-rank constraint) on the learned matrix. For example, Shen et al. [12] obtained a low-dimensional distance metric via the low-rank constraint. Fang et al. [13] proposed an adaptive metric learning with the low-rank constraint method. Unfortunately, imposing the commonly-used low-rank constraint (i.e., the trace-norm of the matrix) on the optimization problem of metric learning still leads to high computational cost. Recently, the Convolutional Neural Network (CNN) based metric learning methods have also been developed [14], [15]. For example, Song et al. [14] proposed a structured clustering loss to train CNN. Wang et al. [15] developed a novel angular loss for deep metric learning, which greatly improves the traditional triplet loss by imposing geometric constraints for triplets. However, these methods suffer from the high computational complexity.

In this paper, we propose a novel Cholesky decomposition based metric learning (CDML) method for effective video-based human action recognition. The similarity matrix is learned by taking advantage of Cholesky decomposition. Compared with the traditional low-rank metric learning methods [12], [13], the proposed method achieves the low-rank property of the similarity matrix by limiting the rank of the lower triangular matrix. Therefore, the proposed method is able to simultaneously perform dimensionality reduction and metric learning, where the optimization problem can be easily and efficiently solved via the gradient descent method. Experimental results show that our method achieves the competitive performance compared with the state-of-the-arts on the challenging video-based human action recognition dataset.

The whole workflow of our algorithm is shown in Figure 1. For each video sequence, the improved Dense Trajectories technique [16] is firstly used for feature detection. For each trajectory, we select three effective feature descriptors for feature extraction. Next, we use the white principal component analysis to reduce the dimensionality and remove noises, and VLDA is employed for feature encoding. Considering that VLDA features have high dimensionality, we further propose to learn a similarity metric to match two sample pairs based on Cholesky decomposition.

Our main contributions of this paper are summarized as follows:

1) The similarity matrix is learned by taking advantage of Cholesky decomposition, which decomposes the matrix into the product between a lower triangular matrix and its symmetric matrix.

2) The proposed method achieves the low-rank property of the similarity matrix by limiting the rank of the lower triangular matrix.

3) Our method is able to simultaneously perform dimensionality reduction and metric learning, where the optimization problem can be easily and efficiently solved via the gradient descent method.
4) The multiple feature descriptors are combined and the improved dense trajectory [16] by removing the background trajectory and the distorted optical flow, is used as the feature detector for human action recognition.

This paper is organized as follows: we firstly describe the feature extraction of action recognition. Secondly, the metric learning algorithm based on Cholesky decomposition is proposed. We then give the experimental results and analysis on public video datasets. Finally, the conclusion is made.

II. FEATURE EXTRACTION OF ACTION RECOGNITION

In this section, we mainly discuss the feature extraction methods for action recognition. As shown in Figure 1, feature extraction is mainly divided into three parts: original low-level feature extraction, feature preprocessing, and feature encoding.

A. THE ORIGINAL LOW-LEVEL FEATURE EXTRACTION

In order to extract the low-level local features, it is generally divided into two steps: Feature detector and feature descriptor. Feature detectors mainly include 3D Harris corner Point [17], Space Time Interest Points [18], Dense Trajectory [19], Improved Dense Trajectory [16], etc. The first two feature detectors are mainly to select the local region which is more advantageous for subsequent recognition and classification by optimizing the loss functions. The latter two feature detectors mainly obtain a large number of points of interest by dense sampling in each frame of the video sequence, and track the detected points of interest to form different dense tracks.

The dense trajectory method [19] is a very classical action recognition algorithm, which deals with dense sampling feature point detector, feature point tracking and feature extraction based on trajectory. Dense trajectory is the points of interest that are intensively collected for each frame in the video sequence. It is generally an extraction of feature points for each scale in multi-scale space (usually 8 scales). The usual scheme of dense trajectory is first to sample an image at W points and track each of these points separately on each scale. Assuming that the sample point in the \( t-th \) frame is \( P_t = (x_t, y_t) \), the position in the \((t+1)-th \) frame through the median filter in the dense optical flow field in \( w = (u_t, v_t) \) is \( P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w)(x_t, y_t) \), where \( M \) is the median filter core; \((x_t, y_t)\) represents the rounding position of \((x_t, y_t)\); and ‘*’ denotes convolution operation. Once the dense optical flow area is detected, the points in the region can be tracked further to form a trajectory, namely \( (P_t, P_{t+1}, P_{t+2}, \ldots) \). In the video-based point tracking, object drifting often occurs. To avoid this problem, you can fix the length of the trace to be \( K \) frames. If the tracked trajectory is longer than \( K \) frames, the feature point is removed. Wang and Schmid [16] further enhances the performance of dense trajectories by estimating camera motion, known as the improved dense trajectory. This method is to improve the performance of the trajectory by using a robust homography matrix to estimate the camera action to remove the background trajectory and the distorted optical flow. In this paper, the improved Dense Trajectories technique is used for feature detection.

In the video-based action recognition algorithms, the feature descriptors include Histogram of Oriented Gradients (HOG) [7], Histogram of Option Flow (HOF) [20] and Action Boundary Histogram (MBH) [21]. Compared with
other feature descriptors, HOG can maintain good invariance to the geometric and optical deformation of the image. Secondly, under the conditions of coarse spatial sampling, fine directional sampling and strong local optical normalization, as long as the pedestrian can maintain the upright posture, HOG can allow the pedestrian to have some subtle body movements, which can be ignored without affecting the detection effect. Therefore, the HOG feature descriptor is particularly suitable for human detection in images. The HOF is used to describe the target’s appearance, which can be characterized by the histogram of each gradient direction in the local region of the image. The basic steps of HOG feature descriptor are: Firstly the target image is divided into several cells and the gradient histogram of each cell is calculated. Then, the multiple cells are grouped into blocks, and all the cells in one block are calculated. The gradient histograms are connected to form the corresponding features of the block structure. Finally, the gradient histograms of the different blocks are connected to obtain the final HOG feature descriptor.

Because the size of the target will change with time, the dimensions of the corresponding optical flow feature descriptors will also change. At the same time, the calculation of optical flow is sensitive to background noise, scale change and motion direction. Therefore, HOF is proposed based on optical flow that can not only represent time-domain change and motion direction. The HOF is mainly used to construct an optical motion information. Therefore, in this paper, we will combine temporal feature descriptor. The HOF describes the optical motion information, which creates a histogram by calculating the action information matrix \(X\). When the feature dimension of the data is high and the correlation between the different features exists, the principal component analysis can make each feature irrelevant in the new projection space, and thus the purposes of both dimensionality reduction and denoising are obtained by preserving the principal components. The specific steps of WPCA are as follows:

1. Calculate the covariance matrix of the training data.

\[
\Sigma = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T.
\]

Here, \(\Sigma\) is called as the covariance matrix of the training data and \(n\) is the number of training samples.

2. The eigenvalues of the covariance matrix \(\Sigma\) and their corresponding eigenvectors are calculated. The eigenvalues are sorted by the descending order to get the eigenvalue set as \(\{\lambda_1, \lambda_2, \ldots, \lambda_d\}\) and the corresponding set of eigenvectors as \(\{v_1, v_2, \ldots, v_d\}\), where \(d\) is the total number of eigenvectors.

3. The eigenvectors corresponding to the largest and the second largest eigenvalues are removed and the eigenvectors corresponding to the first \(p\) eigenvalues \(\Lambda = [\lambda_3, \lambda_4, \ldots, \lambda_{p+2}]\) can be expressed as a linear transformation matrix \(V = [v_3, v_4, \ldots, v_{p+2}]\). In general, the first few dimensions of WPCA correspond to dramatic changes in appearances (such as light, pose, and other external factors). Therefore, this paper proposes to remove the first two-dimensional eigenvector of WPCA to overcome the influence.
of external factors, in order to improve the performance of the subsequent action recognition.

Step 4: For the test sample \( y (y \in \mathbb{R}^d) \), after the feature projection, the reduced dimension feature vector of the sample is \( y' = \text{diag}(\Lambda)V^Ty \). Here, \( V \) is the transformation matrix. \( \text{diag}(\Lambda) \) is the diagonal matrix, and the diagonal elements correspond to the eigenvalues in \( \Lambda \).

The only difference from PCA is that WPCA uses the eigenvalues to whiten the projection features in the final step, whose advantage is that in the new feature space, different features can be normalized to ensure the consistency of different feature scales.

C. FEATURE ENCODING

Feature encoding is often represented by Fisher kernel [22], [23] with local features. The Bag of Words (BOW) is zero moment of Fisher kernel, while the VLAD (Vector of Locally Aggregated Descriptors) feature is the first moment of Fisher kernel.

For the original low-level features extracted from different video sequences, the lengths of eigenvectors obtained are not equal. The length of the eigenvector is related to the number of detected points of interest in the video and the number of frames of the video. The global representation is employed for the original low-level features obtained in each video. Therefore, VLDA is employed for feature encoding in this paper. The feature encoding processing by using VLAD based on Fisher’s kernel representation is depicted in the following.

Step 1: Perform the detection of the points of interest in the sub-sections A and B to extract the local features, and reduce the dimensions using WPCA to obtain the lower-level features.

Step 2: The low-level features are extracted from all the data in the training set, and the codebook is obtained by K-means clustering. At the same time, the points of interest extracted in the video sequence are classified to the nearest clustering center according to the nearest neighbor classification.

Step 3: Calculate the difference of the local feature vectors between the point of interest and the cluster center, then calculate the corresponding histogram, and finally obtain the feature vector with uniform dimension.

III. CHOLESKY DECOMPOSITION BASED METRIC LEARNING ALGORITHM

The traditional similarity metric (or distance metric) methods [31], [32] generally measure the differences between samples by calculating the Euclidean distance, cosine distance or Bhattacharyya distance. However, these similarity metric (or distance metric) methods often treat the different features of the sample equally (without distinguishing the different discriminative abilities among different features), and do not take into account the correlation between different features. Similarity or distance metric learning refers to using a training sample set to obtain a metric model that effectively reflects the similarity (or distance) between the training sample and the test sample. We use the learned similarity metric model (or distance metric model) to measure the similarity (or distance) between the training sample and the test sample, so that the following requirements can be met: The similarity between similar samples becomes larger (or the distance becomes smaller), and the similarity between dissimilar samples becomes smaller (or the distance becomes larger). Typically, these similarity functions can be formulated as \( s(x, y) = x^TMy \), where \( M \) is the similarity matrix, and \( x \) and \( y \) are two feature vectors. By using the learned metric \( s(x, y) \), the similar images become closer to each other, while the dissimilar images become far apart.

In this paper, the image pairs (i.e., similar pairs and dissimilar pairs) are considered for metric learning. To take advantage of the low-rank constraint on the similarity matrix, we first design the following objective function of metric learning:

\[
\min \sum_{(i,j)} l_p(y_{ij}(1 - x_i^TMx_j)) \\
\text{s. t. } \text{rank}(M) \leq d,
\]  

where \( x_i \) and \( x_j \) represent two samples in an image pair. \( y_{ij} \) indicates whether the two samples in the pair are in the same class. Here, \( y_{ij} = 1 \) indicates that the two samples belong to the same class; otherwise they belong to the different classes. \( c \) is the total number of sample pairs. \( l_p(z) \) is a generalized log loss function and \( l_p(z) = 1/\beta \cdot \log(1 + e^{\beta z}) \), where \( \beta \) is the regularization parameter. \( M \) is the similarity matrix to be learned. \( \text{rank}(M) \leq d \) guarantees that the similarity matrix is a low-rank matrix, where \( d \) is a constant to constrain the rank of the similarity matrix.

The low-rank constraint is a typical non-convex optimization problem (an NP-hard problem), which usually leverages the trace norm regularization to achieve the low-rank constraint. Therefore, the original non-convex optimization problem (minimizing the rank of similarity matrix) becomes a convex optimization problem (minimizing the trace norm of similarity matrix). However, the trace norm is often required to iteratively minimize the trace by singular value decomposition, which is not only time-consuming but also easily leads to overfitting (since the solution of the convex problem is only an approximate solution).

This paper uses the Cholesky decomposition, which decomposes the similarity matrix \( M \) into the product between a lower triangular matrix \( U \) and its transpose \( U^T \) (i.e., \( M = UU^T \)). Specifically, based on the Cholesky decomposition of the matrix, we get the following objective function,

\[
\min \sum_{(i,j)} l_p(y_{ij}(1 - x_i^TUU^Tx_j)) \\
\text{s. t. } \text{rank}(U) \leq d.
\]  

In order to constrain the rank of matrix \( U \), we directly set 0 for those column elements whose column numbers exceed.
The solution can be updated as follows:

$$U_{t+1} = U_t - \gamma \nabla U,$$  
(4)

where $t$ is the index number of $t$-th iteration and $\gamma$ is the adjustment parameter for the gradient descent. At the same time, $\nabla U$ represents the gradient, which can be calculated as,

$$\nabla U = \sum_{i,j} -2y_{ij}x_i^T x_j U_{t}(y_{ij}(1-x_iUU^T x_j)).$$  
(5)

In order to constrain the rank of matrix $U$, in each iteration of gradient descent, we directly set 0 for all column elements in the lower triangular matrix whose number of columns exceeds $d$.

The specific steps of the CDML method are shown in Algorithm 1.

**Algorithm 1 Cholesky Decomposition Based Metric Learning (CDML)**

**Input**: data matrix $X$; the similar pairs $S$ and the dissimilar pairs $D$; gradient descent parameter $\gamma$;  
**Output**: mapping matrix $U$;  
1: Initialize $U_0$ by applying WPCA to $X$;  
2: Set $t = 0$;  
3: Repeat:  
4: Calculate the gradient based on (5);  
5: Update the mapping matrix $U_{t+1}$ based on (4);  
6: Set those column elements whose column numbers exceed $d$ to be 0 in the lower triangular matrix;  
7: $t = t + 1$;  
8: Calculate the loss function based on (3);  
9: Until $t$ achieves the maximum iteration number or the convergence condition is satisfied.  
10: Return $U$.

It should be pointed out that, unlike the hint loss function used in the traditional metric learning algorithm, the objective function (as shown in (5)) of the similarity metric learning algorithm proposed in this paper adopts a smoother generalized logical loss function $L_g(z)$. The characteristic of this objective function is that it can be differentiable everywhere, so it is easier to solve the optimization problem using the gradient descent method. At the same time, this paper proposes a new similarity metric algorithm based on Cholesky decomposition of the metric matrix, which converts the metric matrix into a product of a lower triangular matrix $U$ and its transpose $U^T$, which can be controlled by the rank of $U$, thus satisfying the low rank property of the metric matrix.

### IV. EXPERIMENTS

#### A. EXPERIMENTAL SETUP

The experiments are conducted on the ASLAN dataset [24], which is collected in complex natural scenes. It contains 3,697 different videos collected from YouTube, with a total of 432 action categories. The details of the ASLAN dataset are shown in Table 1. Both the training set and the test set use similar pairs and dissimilar pairs. Here, a similar pair refers to actions in two videos that are classified to be the same type of action or behavior. Dissimilar pairs mean that actions in the video are classified as different types of actions. Figure 2 shows some of the similar pairs and dissimilar pairs in the ASLAN dataset. Therefore, the task of video action recognition in this paper is to determine whether the actions occurring in the two video samples are the same type of action.

The performance of video-based action recognition is evaluated by video verification (to decide whether two videos present the same action or not). The verification rate can be used to measure performance for video verification. The rate measures the ratio of the number of the correctly classified pairs and the total number of all pairs. In the ASLAN dataset, we first divide all video sequences into 10 subsets, where each subset contains 350 similar pairs and 350 dissimilar pairs. Then, the performance is evaluated by 10 cross validations. That is, the 9 subsets are randomly selected as the training set, and the rest one is used as the test set. Finally, the mean and variance of the 10 verification rates are taken as the final results.

The methods used for performance comparisons include the Dense Trajectories (DT) method [19], the Low-Rank Metric Learning (LRML) method [12], the Adaptive Metric Learning (ARL) method [13], the Low-Rank Geometric Mean Metric Learning (LR-GMML) method [25], the Multiview Discriminative Analysis of Canonical Correlations (MDACC) method [26], and the multimodal hybrid centroid canonical correlation analysis (MHCCCA) method [27]. The DT, MDACC and MHCCCA methods are the popular and representative video-based action recognition methods. The LRML, LR-GMML and ARL methods are effective similarity (or distance) learning methods based on the low-rank constraint. For all the competing methods, we use the same features as the proposed method for metric learning. In addition, we use the traditional Euclidean distance and Cosine distance as the baselines for comparison.

For the algorithm proposed in this paper, we use the improved dense trajectory [16] as the feature detector in

### TABLE 1. Details of the ASLAN dataset.

|                         |                  |
|-------------------------|------------------|
| Total number of action categories | 432              |
| Number of action samples | 3,697            |
| Number of action categories with > 1 sample | 316              |
| Number of long action samples (> 10 sec) | 71               |
| Number of long action samples (< 1 sec)    | 187              |
the experiment. First, for each trajectory in the dense trajectory, we select three feature descriptors, i.e., HOG, HOF and MBH. The HOG feature descriptor is 96-dimensional, the HOF feature descriptor is 108-dimensional, and the MBH feature descriptor is 192-dimensional. Finally, the low-level local features of 396 dimensions are obtained. Next, WPCA is used to reduce and denoise the low-level local features, and the 396 dimension is reduced to the 20-dimensional feature space. Then, the VLAD features are used for encoding, and the number of the codebook centers of clustering are 256, so the features obtained after the VLAD encoding process are 5,120 dimensions. Finally, a similarity metric learning algorithm based on Cholesky decomposition is used to learn a metric. In addition, we use the traditional cosine distance and Euclidean distance as the benchmark (using the same features as the algorithm in this paper) for comparisons. All the competing methods are implemented using MATLAB with an Intel 6700K 4.0 GHz CPU and 8.00 GB RAM.

B. EXPERIMENTAL RESULTS AND ANALYSIS
In experiments, we respectively use a single feature descriptor and the combined feature descriptor for feature extraction. The experimental results using a single feature descriptor are shown in Table 2. In addition, the results obtained by the different competing methods (using the combined feature descriptor) are shown in Figure 3.

From Table 2, we can see that based on a single feature descriptor, compared with the Cosine distance and the Euclidean distance, the similarity learning (or distance learning) methods are beneficial to improve the recognition rate (about 3% to 7% improvements), which shows the effectiveness and importance of metric learning. The results obtained by the Cosine distance are slightly better than those obtained by the Euclidean distance. This is because the Cosine distance measures the changes in angle (removing the effect of different scales on the feature vectors). The proposed method obtains about 7% improvements in the verification rate for each feature descriptor, compared with the Cosine distance. Compared with LRML, ARL and LR-GMML, our CDML method obtains the 2%~5% improvements in performance. Moreover, the proposed CDML method requires the less training time than the other competing methods except for the DT method, but the higher video verification rates are obtained by our method than the DT method. Therefore, our method can effectively learn the similarity between different video actions, since the low-rank matrix based on Cholesky...
decomposition not only effectively achieves dimensionality reduction, but also contains discriminative information that can be used to identify different video actions.

Figure 3 shows the video verification rates obtained by the different methods by employing the combined feature descriptor (i.e., HOG+HOF+MBH). It can be seen that using metric learning can effectively improve the performance in video-based action recognition. Compared with a single feature descriptor, the combined feature descriptor can further improve the performance of all methods. As shown in Table 2, based on a single MBH feature descriptor, the mean video verification rates of our method is highest, i.e., 60.59%, and however, it is lower than the video verification rate of 63.3% obtained by the proposed algorithm on the combined feature descriptor. This is mainly because the three feature descriptors describe the information of the video sequence from different aspects: The HOG mainly describes the gradient information in the local area of the image. The HOF describes the optical flow gradient, while the MBH characterizes the motion information of the foreground target. Therefore, the proposed method considers the advantages of the HOG, HOF and MBH feature descriptors, and the effectiveness of the combined feature descriptor is verified. In addition, compared with the two representative video-based action recognition methods (i.e., MDACC and MHCCCA), the CDML method obtains the better results, due to the effective metric learning based on Cholesky decomposition.

In particular, compared with the LRML algorithm and the ARL algorithm, we can see that the proposed similarity metric learning algorithm can effectively distinguish the changes of the actions in the different video sequences.

V. CONCLUSION

This paper proposes a new metric learning method called CDML based on Cholesky decomposition for video-based human action recognition. Different from the traditional low-rank metric learning methods, our method takes advantage of Cholesky decomposition to decompose the similarity matrix into the product between a lower triangular matrix and its symmetric matrix. In this way, the low-rank constraint of the matrix can be naturally achieved by controlling the rank of the lower triangular matrix. Therefore, to avoid the problems existing in the traditional methods that require the repeatedly iterative calculations of the singular value decomposition, the proposed method effectively uses the learned matrix based on Cholesky decomposition, where the samples in the high-dimensional feature space can not only be projected onto the low-dimensional feature space, but also the similarity between similar samples is larger and meanwhile the similarity between dissimilar samples is smaller. Experimental results show the effectiveness and efficiency of the proposed method. In future, we will further evaluate the proposed algorithm in the real applications.

REFERENCES

[1] Y. Wang, S. Cang, and H. Yu, “A survey on wearable sensor modality centred human activity recognition in health care,” Expert Syst. Appl., vol. 137, pp. 167–190, Dec. 2019.
[2] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu, “Gaze-informed egocentric action recognition for memory aid systems,” IEEE Access, vol. 6, pp. 12894–12904, Mar. 2018.
[3] J. Ng, J. Choi, J. Neumann, and L. Davis, “ActionFlowNet: Learning motion representation for action recognition,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Lake Tahoe, NV, USA, Mar. 2018, pp. 1616–1624.
[4] M. Subramani, B. Lubomir, and M. Jitendra, “Action recognition from a distributed representation of pose and appearance,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Colorado Springs, CO, USA, Jun. 2011, pp. 3177–3184.
[5] S. Maji, L. Bourdev, and J. Malik, “Action recognition from weak alignment of body parts,” in Proc. Brit. Mach. Vis. Conf. (BMVC), Nottingham, U.K., 2014, pp. 1–12.
[6] B. Yao, A. Khosla, and F. Li, “Combining randomization discrimination for fine-grained image categorization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Colorado Springs, CO, USA, Jun. 2011, pp. 1577–1584.
[7] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), San Diego, CA, USA, Jul. 2005, pp. 886–893.
[8] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.
[9] G. Gkioxari, R. Girshick, and J. Malik, “Contextual action recognition with R*CNN,” in Proc. IEEE Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1080–1088.
[10] X. Peng and C. Schmid, “Multi-region two-stream R-CNN for action detection,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, 2016, pp. 744–759.
[11] Y. Wang, L. Zhou, and Y. Qiao, “Temporal hallucinating for action recognition with few still images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 1–8.
[12] Y. Shen, Y. Yan, and H. Wang, “Low rank metric learning via accelerated gradient optimization,” in Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS), Xiamen, China, 2014, pp. 299–302.

[13] Y. Fang, Y. Yan, H. Wang, S. Chen, and X. Gao, “Adaptive metric learning with the low rank constraint,” in Proc. Int. Conf. Internet Multimedia Comput. Service (ICIMCS), Xi’an, China, 2016, pp. 61–65.

[14] H. Song, S. Jegelka, and V. Rathod, “Deep metric learning via facility location,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 5382–5390.

[15] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, “Deep metric learning with angular loss,” in Proc. Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2593–2601.

[16] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in Proc. Int. Conf. Comput. Vis. (ICCV), Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.

[17] I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp, “A robust 3D interest points detector based on Harris operator,” in Proc. Eurograph. Workshop 3D Object Recog., Norrköping, Sweden, 2010, pp. 1–7.

[18] I. Laptev, “On space-time interest points,” Int. J. Comput. Vis., vol. 64, nos. 2–3, pp. 107–123, Sep. 2005.

[19] H. Wang, A. Klaser, C. Schmid, and C. Liu, “Action recognition by dense trajectories,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Colorado Springs, CO, USA, Jul. 2011, pp. 3169–3176.

[20] A. Efros, A. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” in Proc. IEEE Conf. Comput. Vis. (ICCV), Nice, France, Oct. 2003, pp. 726–733.

[21] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Berlin, Germany: Springer, 2006, pp. 428–441.

[22] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in Proc. Int. Conf. Comput. Vis. (ICCV), Rio de Janeiro, Brazil, Jan. 2007, pp. 1–8.

[23] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, “Aggregating local image descriptors into compact codes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 9, pp. 1704–1716, Dec. 2011.

[24] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[25] M. Bhattu, P. Jawanpura, H. Kasai, and B. Mishra, “Low-rank geometric mean metric learning,” 2018, arXiv:1806.05454. [Online]. Available: https://arxiv.org/abs/1806.05454

[26] N. E. Madany, Y. He, and L. Guan, “Human action recognition via multiview discriminative analysis of canonical correlations,” in Proc. IEEE Conf. Image Process. (ICIP), Seoul, South Korea, Sep. 2016, pp. 4170–4174.

[27] N. E. Madany, Y. He, and L. Guan, “Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis,” IEEE Trans. Multimedia, vol. 21, no. 5, pp. 1317–1331, May 2018.

[28] E. Chen, X. Bui, L. Gao, H. C. Tinega, and Y. Ding, “A spatiotemporal heterogeneous two-stream network for action recognition,” IEEE Access, vol. 7, pp. 57267–57275, May 2019.

[29] W. Min, H. Cui, H. Rao, Z. Li, and L. Yao, “Detection of human falls on furniture using scene analysis based on deep learning and activity characteristics,” IEEE Access, vol. 6, pp. 9324–9335, Mar. 2018.

[30] Y. Tang, J. Lu, Z. Wang, M. Yang, and J. Zhou, “Learning semantics-preserving attention and contextual interaction for group activity recognition,” IEEE Access, vol. 28, pp. 4997–5012, Oct. 2019.

[31] J. Hu, J. Lu, and Y.-P. Tan, “Sharable and individual multi-view metric learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 9, pp. 2281–2288, Sep. 2018.

[32] D. Tran and A. Sorokin, “Human activity recognition with metric learning,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Berlin, Germany: Springer, 2008, pp. 548–561.

[33] H. Liu, J. Lu, J. Feng, and J. Zhou, “Two-stream transformer networks for video-based face alignment,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 11, pp. 2546–2554, Nov. 2018.