View-Invariant Feature Discovering for Multi-Camera Human Action Recognition

Hong Lin¹², Lekha Chaisorn², Yongkang Wong², An-An Liu¹, Yu-Ting Su¹, Mohan S Kankanhalli²³

¹School of Electronic Information Engineering, Tianjin University, Tianjin, China
²SeSaMe Center, Interactive & Digital Media Institute, National University of Singapore, Singapore
³School of Computing, National University of Singapore, Singapore
linhong080324@gmail.com, {idmlekh, yongkang.wong}@nus.edu.sg, anan0422@gmail.com, ytsu@tju.edu.cn, mohan@comp.nus.edu.sg

Abstract—Intelligent video surveillance system is built to automatically detect events of interest, especially on object tracking and behavior understanding. In this paper, we focus on the task of human action recognition under surveillance environment, specifically in a multi-camera monitoring scene. Despite many approaches have achieved success in recognizing human action from video sequences, they are designed for single view and generally not robust against viewpoint invariant. Human action recognition across different views remains challenging due to the large variations from one view to another. We present a framework to solve the problem of transferring action models learned in one view (source view) to another view (target view). First, local space-time interest point feature and global shape-flow feature are extracted as low-level feature, followed by building the hybrid Bag-of-Words model for each action sequence. The data distribution of relevant actions from source view and target view are linked via a cross-view discriminative dictionary learning method. Through the view-adaptive dictionary pair learned by the method, the data from source and target view can be respectively mapped into a common space which is viewpoint invariant. Furthermore, We extend our framework to transfer action models from multiple views to one view when there are multiple source views available. Experiments on the IXMAS human action dataset, which contains videos captured with five viewpoints, show the efficacy of our framework.

I. I N T R O D U C T I O N

Intelligent visual surveillance system is directed on automatically identifying events of interest. It has obtained increasing attention due to its effective function in security surveillance, such as production monitoring and unusual event alarming. Most of the events of interest, tend to involve object or human (e.g., [1], [2], [3]) as the center of attention. With increasingly equipped surveillance cameras, an intelligent visual surveillance system is very much needful in private and public locations due to recognize and prevent unusual or suspicious activities. For example, it can play a major role in home health monitoring for the elderly aging at home. Also it’s helpful for observing people within a busy environment to detecting illegal actions and events such as shopping mall, hospital and airport. A general framework of human action recognition in visual surveillance system is shown in Fig. 1. The processing steps include feature extraction, action recognition, and alarm system. Feature extraction, where representative descriptors are obtained at each time interval. The choice of features that make up the feature vector is an important design decision in the action classification part of problem. Action recognition, where the features of the data and the way they change with time are analyzed and used to classify it. Various classifiers have been used to recognize the event. The choice of the classifiers used is highly dependent on the feature set. Finally, if unusual activity is detected, the system will alert the human operator to take action against it.

Recently, many appearance-based approaches for action recognition in video sequences are proposed, such as optical flow patterns [4], learned geometrical models of the human body parts [5], shape features [6], and spatio-temporal interest point [7]. Although these approaches are quite successful in recognizing actions captured from similar views, they have more limitation when the viewpoint changes. The main reason lies in the appearance of actions may be drastically different
when observed from different views. Consequently, the action models learned using low-level features become less discriminative. However, in the actual monitoring environment, especially large public place, the area is observed by different views at the same time. More cameras mean more monitor terminals and view data, and they bring more computational cost to visual surveillance system. So we more concern about cross-view human action recognition in multi-camera scene. Also, how to take advantage of multiple views information is another key point.

Motivated by these aspects, we focus on the method to realize cross-view human action recognition in visual surveillance system. Our work is devoted to discover a more discriminative feature representation based on the low-level feature representation. This view-invariant feature representation bridges different feature spaces due to different views data. Based on it, the source view action classifiers can be directly used for action recognition in the target view. Thus, we reach our objective of cross-view human action recognition. Furthermore, we fuse multiple views data together to transfer action models learned from multiple source views to one target view, as an extention.

The contributions of our work are two-fold:

- We realize an effective cross-view action recognition method which discovers view-invariant feature representations based on view-specific low-level feature representations.
- We extend the method to utilize multi-view knowledge when there are multiple source views available.

The paper is organized as follows. In Section II, we discuss on the related work. Then, we introduce our proposed method in Section III. Experimental results and comparisons with state-of-the-art approaches are shown in Section IV. Section V concludes the paper.

II. RELATED WORKS

This paper focuses on cross-view action recognition in video sequences, a topic which has received a considerable attention from researchers recently. To address this problem, several geometry-based approaches was employed for this task. Yilmaz et al. [8] employed epipolar geometry for point correspondences between actions to impose fundamental matrix constrains for view-invariant action recognition. Rao et al. [9] showed that the maxima in space-time curvature of a 3D trajectory were preserved in 2D image trajectories, and therefore the 2D trajectories can be utilized to capture the view-independent representation. Another kind of approaches perform 3D reconstruction to take multi-view information for multi-view action recognition. Lv et al. [10] constructed the Action Net model to represent 3D shapes for action recognition. Li et al. [11] reconstructed 3D model from multi-view inputs for action recognition. However, 3D recognition usually requires pre-setup of multi-view cameras to get strict alignment between views. Also, the computationally expense limits their application in practice. Rather than using view-based knowledge for action recognition, Junejo et al. [12] employed a very simple and interesting action representation called Self-Similarity Matrix that can capture the structure of temporal similarities and dissimilarities with an action sequence.

The approaches most closely related to our approach are that of [13], [14], [15], [16]. These transfer learning based methods have emerged to adapt the action knowledge learned on one or more views (source view) to another different view (target view) by exploring the statistical connections between them. Farhadi et al. [13] proposed a split-based feature which learned by Maximum Margin Clustering algorithm. This split-based feature first generated in one view, then a predictor was used to obtain the split-based feature in the other view for action recognition. Their work requires feature-to-feature correspondence at the frame-level to train a classifier to obtain split-based features. Liu et al. [14] employed a bipartite graph-based method to learn bilingual-words from two individual codebooks belong to two different views and then transferred action models between two views by a bag-of-bilingual-words (BoW) model. However, this method only exploits the codebook-to-codebook correspondence in two views, which can not ensure pairs of videos taken at two views have similar feature representation. Li et al. [15] proposed “virtual views” to connect action descriptors between different views. Each virtual view is associated with a linear transformation of the action descriptor, and the sequence of transformed descriptors can be used to compare actions from different views. The most encouraging method is proposed by Zheng et al. [16], who learn two dictionaries of source and target views simultaneously to ensure same action to have same representation. Although this algorithm achieve good performance overall, it is harder to transfer action models when involves the top view.

III. PROPOSED METHOD

A. Overview

In multi-camera scene, multiple views information is available since the area is observed by different cameras. The objective of our work is to recognize an unknown action from one view (target view) using training data taken from another view (source view). We first extract space-time interest point feature from both source and target view action videos using Harris 3D detector [7] and Histogram of Oriented Gradients (HOG)/Histogram of Optical Flow (HOF) descriptor [17]. Given the extracted features, we construct codebooks for each view respectively using k-means clustering algorithm. Then each action video can be further models as Bag-of-Words (BoW) [18] using the corresponding codebook. As a complement to the space-time interest point feature, we extract motion context feature [19] from each action video. The motion context feature is then used to build BoW model the same way as the space-time interest point feature. Then, for each action video, we concatenate space-time interest point feature based BoW model and motion context feature based BoW model to form a hybrid model. Since each view has its own codebook, such representation of action videos in the two views are belong to two different feature spaces. Based on
B. Low-Level Feature Extraction

The low-level feature used in our framework is the local space-time interest point feature and the global motion context feature. The interest point feature can capture local salient characteristics of appearance and motion as well as being tolerant to occlusion, background clutter, and scale changes. But it is not able to capture the global shape information. So we further extract shape and flow feature, called motion context descriptor, as proposed in [19]. We regard it as a complementary to the interest point feature, motivated by the work in [14]. We give a brief review of these two features in Section III-B1 and III-B2 respectively.

1) Local Space-Time Interest Point Feature: Local space-time interest point feature is usually extracted directly from video and therefore avoid possible failures of other preprocessing methods such as motion segmentation and tracking. We detect the space-time interest points using the Harris 3D detector [7]. A spatio-temporal second-moment matrix at each video point \( \mu(\sigma, \tau) = \varphi(-s, s, \tau) * (\nabla L(\varphi(\sigma, \tau) \nabla L(\varphi(\sigma, \tau)))^T) \) using independent spatial \( \sigma \) and temporal scale values \( \tau \), a separable Gaussian smoothing function \( \varphi \), and space-time gradients \( \nabla L \). The final locations of space-time interest points are given by local maxima of \( H = det(\mu) - n trace^2(\mu) \), which ensure \( H > 0 \). Interest points detected for some action videos according to different views of IXMAS dataset are illustrated in Fig. 3. We can see that the detected points correspond to the strong spatio-temporal variation of the image data. Given the the detected points, HOG/HOF descriptor [17] is used to compute histograms of spatial gradient and optic flow accumulated in space-time neighborhoods of them.

2) Global Motion Context Feature: Motion context descriptor is a histogram of the silhouette and of the optic flow inside the normalized bounding box. The optic flow measurements are split into horizontal and vertical channels. So, the three information channels, namely horizontal optical flow, vertical optical flow, and silhouette, are extracted from every frame. The Principal Component Analysis (PCA) algorithm is used to reduce the dimension of all these features. Furthermore, in order to capture temporal information, the feature information from neighbor frames are integrated into the current frame descriptor by simply concatenating feature vectors. We refer the reader to [19] for more details of motion context feature.

C. Hybrid BoW Feature Modelling

Given the space-time interest point feature and motion context feature extracted from action videos, we build a hybrid BoW model for each action video. For space-time interest in IXMAS. The first row to the fifth row show the detected interest points corresponding to action videos from Camera 0 to Camera 4.

Fig. 2: The flowchart to discover view-invariant feature representation

Fig. 3: Space-time interest points detected for some action videos
point feature, we construct codebook for each view separately which consists of a set of visual-words by k-means clustering algorithm. Then, each video is modelled as a BoW using the corresponding codebook. Follow the same way, we build BoW model for each action video based on motion context feature. Finally, for each action video, we concatenate this two models to build a hybrid BoW model. We believe this hybrid BoW model could make a better description of action videos.

D. View-invariant Feature Representation

In this section, we present the algorithm to obtain view-invariant feature representation based on the K-Singular Value Decomposition (K-SVD) algorithm [20]. It generalizes the k-means clustering process for adapting dictionaries to efficiently learn an over-complete dictionary from a set of training signals. Let \( Y \) be a set of \( n \)-dimensional input signals which contains \( N \) instances, i.e., \( Y = [y_1, \ldots, y_N] \in \mathbb{R}^{n \times N} \). Learning a reconstructive dictionary for getting the sparse representation of \( Y \) can be accomplished by solving the following optimization problem:

\[
\langle D, X \rangle = \arg \min_{D, X} \| Y - DX \|_2^2 \quad \text{s.t.} \forall i, \| x_i \|_0 \leq T
\]

(1)

where \( D = [d_1, \ldots, d_k] \in \mathbb{R}^{n \times k} \) is the learned dictionary, \( X = [x_1, \ldots, x_N] \in \mathbb{R}^{k \times N} \) are the sparse representation of \( Y \). \( T \) is a sparsity constraint factor, and \( \| x_i \| \leq T \) means each signal has fewer than \( T \) items in its decomposition. The term \( \| Y - DX \|_2^2 \) denotes the reconstruction error. The construction of \( D \) is achieved by minimizing the reconstruction error and satisfying the sparsity constraints. The K-SVD algorithm is an iterative approach to solve Eq 1 and learns a reconstructive dictionary for sparse representations of signals. Assuming a dictionary \( D \) is given, the sparse representation \( X \) of \( Y \) can be obtained by solving Eq 2 using the orthogonal matching pursuit (OMP) [21] algorithm.

\[
X = \arg \min_X \| Y - DX \|_2^2 \quad \text{s.t.} \forall i, \| x_i \|_0 \leq T
\]

(2)

1) View-invariant Feature Discovery from Pairwise Views:

In a pair of views, the two sets of videos are also paired. Let \( Y_s, Y_t \) denote the feature representation of action videos in source and target view individually. Assuming instances in \( Y_s = [y_{s,1}, \ldots, y_{s,n}] \) and \( Y_t = [y_{t,1}, \ldots, y_{t,n}] \) are arranged according to their pair relation. In these two views, the optimization problem for dictionary learning are:

\[
\langle D_s, X_s \rangle = \arg \min_{D_s, X_s} \| Y_s - D_sX_s \|_2^2 \quad \text{s.t.} \forall i, \| x_{s,i} \|_0 \leq T
\]

(3)

\[
\langle D_t, X_t \rangle = \arg \min_{D_t, X_t} \| Y_t - D_tX_t \|_2^2 \quad \text{s.t.} \forall i, \| x_{t,i} \|_0 \leq T
\]

(4)

By minimizing the reconstruction error terms \( \| Y_s - D_sX_s \|_2^2 \) and \( \| Y_t - D_tX_t \|_2^2 \) in Eq 3 and Eq 4, respectively, there is no doubt that the sparse representation \( X_s \) and \( X_t \) still obey to the respective feature space. In order to force the different representations from different views to the common feature space, we combine the objective functions in Eq 3 and Eq 4 to build a new function:

\[
\langle D_s, D_t, X_s, X_t \rangle = \arg \min_{D_s, D_t, X_s, X_t} \| Y_s - D_sX_s \|_2^2 + \| Y_t - D_tX_t \|_2^2 \quad \text{s.t.} \forall i, \| x_{s,i} \|_0, \| x_{t,i} \|_0 \leq T
\]

(5)

Given the same number of action videos and same arrangement in source and target view, \( D_s \) and \( D_t \) have the same number of dictionary items. At the same time, we force \( X_s \) and \( X_t \) to be the same \( X \), such that \( D_s \) and \( D_t \) encode the ability to map two different feature spaces to a common one. Eq 5 can be rewritten as:

\[
\langle D_s, D_t, X \rangle = \arg \min_{D_s, D_t, X} \| Y_s - D_sX \|_2^2 + \| Y_t - D_tX \|_2^2 \quad \text{s.t.} \forall i, \| x_{i} \|_0 \leq T
\]

(6)

Thus, the dictionary pair \( D_s \) and \( D_t \) can be learned through the K-SVD algorithm. After getting the dictionary pair \( D_s \) and \( D_t \), the view-invariant feature representations \( X_s \) and \( X_t \) can be obtained by the OMP algorithm. Based on such view-invariant feature representation, a classifier learned in the source view can be used to recognize the testing videos in the target view.

2) View-invariant Feature Discovery from Multiple Views:

Suppose there are more than one source views, one problem is how to explore the benefits of combining the multiple views knowledge to improve the recognition performance in the target view. As an extension of view-invariant feature discovery from pairwise views, we learn a set of dictionaries by forcing action videos of all views to have same representations. The set of dictionaries can be obtained by solving:

\[
\langle D_{s,1}, \ldots, D_{s,n}, D_t, X \rangle = \arg \min_{\{D_{s,i}\}_{i=1}^n, D_t, X} \sum_{i=1}^n \| Y_{s,i} - D_{s,i}X \|_2^2 + \| Y_t - D_tX \|_2^2 \quad \text{s.t.} \forall i, \| x_{i} \|_0 \leq T
\]

(7)

Similarly, Eq 7 can be rewritten as:

\[
\langle D_{s,1}, \ldots, D_{s,n}, D_t, X \rangle = \arg \min_{\{D_{s,i}\}_{i=1}^n, D_t, X} \left\| \begin{pmatrix} Y_{s,1} \\ \cdots \\ Y_{s,n} \end{pmatrix} - \begin{pmatrix} D_{s,1} \\ \cdots \\ D_{s,n} \end{pmatrix} X \right\|_2^2 \quad \text{s.t.} \forall i, \| x_{i} \|_0 \leq T
\]

(8)

Now, with the dictionaries set \( \{D_{s,1}, D_{s,2}, \ldots, D_{s,n}, D_t\} \), which corresponding to \( n \) source views and one target view,
we can map each action videos into a common view-invariant feature space using the corresponding view-dependent dictionary. Thus, we do not need to differentiate the training videos from each source view in this view-invariant feature space. Any action model learned using all the training videos in all source views can be directly used to recognize the testing videos in the target view.

IV. EXPERIMENTS

A. Dataset

We evaluate our framework on IXMAS multi-view dataset [22] which contains 11 daily-live actions\(^1\). Each action was performed 3 times by 10 actors and simultaneously recorded with 5 cameras. Each video sequence has the frame rate of 23 frames per second and the frame size of 390 × 291 pixels. The actors freely chosen position and orientation for each sequence. Example snapshots of the IXMAS dataset are shown in Fig. 4.

B. Implementation Details

We follow the method described in [17] to extract local space-time features and used the code released by the authors. The code is an extension of Harris 3D detector proposed in [7] which detects space-time points of interest. HOG/HOF descriptor are then computed to characterize the 3D space-time video patch extracted at each interest point. The 3D video patch is divided into \(3 \times 3 \times 2\) cells where 4-bin HOG and 5-bin HOF are computed in each cell. All the feature, 72-dimension HOG feature and 90-dimension HOF feature, are concatenated and normalized to form a single histogram. Note that the code does not implement scale selection as in [7], instead interest points are detected at multiple levels of spatial and temporal scales \((\sigma^2, \tau^2)\). We use the standard parameter setting \(k = 0.0005, \sigma^2 = \{4, 8, 16, 32, 64, 128\}, \tau^2 = \{2, 4\}\). For the motion context feature, we use the code available online\(^2\) and detect features using original implementation. Three 72-dimensional histograms corresponding to horizontal optical flow, vertical optical flow and silhouette are concatenated to obtain a 216-dimensional descriptor. Then, 70-dimensional descriptor of temporal information is appended to form the final 286-dimensional motion context descriptor. For classification, we employ \(k\)-NN classifier to predict the testing videos in the target view. We traverse \(k\) from 1 to 10 to find the best performance in our experiments.

We compare our method with two state-of-the-art unsupervised cross-view action recognition approaches reported in [14] and [16]. We follow the same leave one action out strategy in our evaluation. In other words, we select one action class for testing in the target view for each test phase, while this action class is not used to learn the dictionary pair in both source and target views. The average accuracy for all action classes in each view are reported. In total, there are 20 sets of cross-view experiment.

C. Results on Pairwise Cross-view Action Recognition

In this section, we evaluate our framework for pairwise cross-view action recognition on IXMAS dataset. We first extract local space-time interest feature from action videos of five different views. With five sets of local space-time features corresponding to five different views, we learn five codebooks of 1000 words by \(k\)-means clustering for five views individually. Each action video is represented by 1000-dimensional BoW model using the corresponding view-dependent codebook. As mentioned in Section III-B2, we extract motion context features for global information. We learn five codebook of 500 words of motion context descriptor. Then each video in each view is modeled similar to local feature. Finally, each action video in each view has a 1500-dimensional hybrid BoW model.

The experimental results are presented in Table I. We note that our performance is much better than [14] in all cases. For the most encouraging unsupervised approach in [16], it’s interesting to find that our approach performs better than [16] when the Cam 2 or Cam 4 is the source or target view. Although a part of our recognition accuracy are slightly lower than [16], the overall accuracy of our approach is 93.0%, which is about 3% higher than that of [16]. Since the Cam 4 observed actors in the top view, we believe that the recognition accuracy of Cam 4 are more significant for evaluating a cross-view action recognition approach. The excellent performance in Cam 4 demonstrates the efficacy of our framework.

<table>
<thead>
<tr>
<th>Cam 0</th>
<th>Cam 1</th>
<th>Cam 2</th>
<th>Cam 3</th>
<th>Cam 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cam 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cam 1</td>
<td>79.9</td>
<td>96.7</td>
<td>96.1</td>
<td>76.8</td>
</tr>
<tr>
<td>Cam 2</td>
<td>79.6</td>
<td>92.1</td>
<td>96.1</td>
<td>76.6</td>
</tr>
<tr>
<td>Cam 3</td>
<td>73.0</td>
<td>97.0</td>
<td>91.8</td>
<td>74.1</td>
</tr>
<tr>
<td>Cam 4</td>
<td>82.0</td>
<td>83.0</td>
<td>98.5</td>
<td>68.3</td>
</tr>
<tr>
<td>Ave.</td>
<td>79.0</td>
<td>92.4</td>
<td>94.6</td>
<td>74.7</td>
</tr>
</tbody>
</table>

\(^1\)check watch, cross arm, scratch head, sit down, get up, turn around, walk, wave, punch, kick and pick up

\(^2\)http://visionpc.cs.uiuc.edu/projects/activity/
D. Results on Multiple Views Action Recognition

In the above experiments, we evaluate the performance of pairwise cross-view action recognition. For each view, there are four recognition results obtained from the classification models trained on the rest four views. In this section, we use the method mentioned in Section III-D2 to utilize multiple views knowledge aim to making the classification model more robust. Action models are learned using all training videos in all source views and then directly used to classify the testing videos in the target view. The results are shown in Table II. Overall, our performance is better than [14] and competitive to [16]. Especially, we observe that our proposed method achieves a nearly perfect recognition accuracy on the top view (Cam 4), and outperforms [16]’s unsupervised approaches as well. This also demonstrates the effective of our framework, especially for top view.

V. CONCLUSION

In this paper, we realize a cross-view action recognition framework which is effective and easy to follow. Our objective is to recognize an unknown action from one (target) view using action models learned from other (source) views. For this purpose, we first extract spatio-temporal interest point feature (Harris 3D detector and HOG/HOF descriptor) and global shape-flow feature as low-level feature. Then, each action video of each view is modeled as hybrid Bag-of-Words (BoW) model with corresponding view-specific codebook. By forcing two sets of videos from different views to have the same representation, we simultaneously learn the corresponding dictionary pair. Using the learned dictionary pair, the action videos from the two views can be represented with view-invariant feature. Therefore, we can directly transfer action models across views. Furthermore, we extend the framework to take advantage of multiple views information to improve recognition performance in target view. Experimental results show the efficacy of our framework.

VI. ACKNOWLEDGEMENT

This research was carried out at the NUS-ZJU Sensor-Enhanced Social Media (SeSaMe) Centre. It is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office. The work is partially supported by the National Natural Science Foundation of China (61100124, 21106095, 61170239, 61202168) and the grant of Elite Scholar Program of Tianjin University.

TABLE II: Multi-view action recognition results (in percentage) of the proposed method and various approaches.

<table>
<thead>
<tr>
<th></th>
<th>Cam 0</th>
<th>Cam 1</th>
<th>Cam 2</th>
<th>Cam 3</th>
<th>Cam 4</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>99.4</td>
<td>100.0</td>
<td>97.6</td>
<td>99.1</td>
<td>96.9</td>
<td>98.6</td>
</tr>
<tr>
<td>Zheng et al. [16]</td>
<td>98.5</td>
<td>99.1</td>
<td>99.1</td>
<td>100.0</td>
<td>90.3</td>
<td>97.4</td>
</tr>
<tr>
<td>Liu et al. [14]</td>
<td>76.7</td>
<td>73.3</td>
<td>72.9</td>
<td>73.0</td>
<td>N/A</td>
<td>73.8</td>
</tr>
</tbody>
</table>

REFERENCES