Human Behaviour Analysis based on Sparse Coding

Tonghuan Hua^{1,*}, Xiaowei Li², Dan Chen³, Qiang Wang⁴

¹Hubei College of Traditional Chinese Medicine, Hubei, China

²Red Hat Software, Beijing, China

³IBM, China

⁴CSR Zhuzhou Institute CO.,LTD, China

*Corresponding Author: tonghuan.hua@gmail.com

Copyright ©2014 Horizon Research Publishing All rights reserved.

Abstract Sparse coding and compressive sensing have attracted lots of interest in the computer vision area. This paper proposes a new scheme to recognize human motions in video sequences based on the sparse representation of image frames. Each frame of a video is transformed to a linear combination of a few elements in a dictionary. The class label of the video is determined based on the reconstruction errors of individual frames or the overall reconstruction error of the video. A series of experiments were conducted to evaluate the performance of the proposed method. Experimental results demonstrate that the sparse representation method achieves an accuracy on par with or exceeding that of existing methods.

Keywords ℓ_1 norm, behaviour recognition, sparse coding

sparse coding to obtain an accurate and discriminative representation. However, our method is different from it in that image frames of a video are encoded while the work in [9] encodes local spatial-temporal patches. Moreover, our method makes the classification via minimum of reconstruction errors of frames, thus does not need a further classifier.

The main contribution of the work is two-fold. Firstly, we proposed a new scheme based on sparse coding to recognize human behaviour in videos. In addition, we extended the sparse representation classification method (SRC) for an individual image to classify a video that contain a set of frames by introducing an overall reconstruction error.

The remainder of the paper is organized as follows. Section 2 briefly describes the feature representation for image frames. Section 3 gives the details of the sparse representation for classification of video sequences. Section 4 analyzes the experimental results of our method on a public dataset. Finally, conclusion is given in Section 5.

1 Introduction

Automatic human behaviour recognition in video has been widely investigated in the community of computer vision due to its various applications in security surveillance, human-computer interaction, and animation [1]. Some works extract 3D spatial-temporal interest points as features [2][3], while some others use information of frames to characterize a video [4]. After features are extracted, generative [3] or discriminative [2][5] classifiers are used to determine the class of a video.

Recently, sparse coding becomes a hot topic in computer vision field. It has been used for image inpainting [6], face recognition [7], object classification [8] and etc. In this paper, we proposed a new scheme for behaviour classification based on sparse coding. Image frames in a video are transformed to a linear combination of a few atoms in a dictionary. Reconstruction errors of the frames from sparse representations are employed to make the final decision. The work in [9] also applies the

2 Frame representation

We use the Pyramid Histograms of Gradient Orientations (Pyramid HoG) descriptor proposed in [10] to represent video frames. The basic idea is to divide a tracked and localized interest area into a number of cells at several pyramid levels. Gradient orientation on all pixels within each cell is accumulated to form a histogram. The final representation is constructed by concatenating all the histogram.

Assuming that subjects in video sequences are tracked and localized, we normalize all interest regions to a fix size $M \times N$ (i.e., $M = 96, N = 64$) so that they are at the same spatial scale. After normalized, each interesting region is divided into small spatial cells into different pyramid levels. Then, gradients over all the pixels within a cell are accumulated to form a local bins histogram. Finally, all the cells at different pyramid lev-

els are concatenated to form a final representation. This representation is then projected to a lower dimensional space by the principal component analysis (PCA). Readers are recommended to refer [10] for more details about the Pyramid HoG descriptor.

3 Classification via Sparse Representation

3.1 Sparse coding on image frames

Suppose that there are M videos in training data, each video j contains T_j frames, we extract the pyramid HoG descriptor with a dimension of d for each of the T_j frames. Then a video is represented as a set of features $\mathbf{X}_j = [\mathbf{x}_j^1, \mathbf{x}_j^2, \dots, \mathbf{x}_j^{T_j}] \in \mathbb{R}^{d \times T_j}$. Given an over-complete dictionary $\mathbf{V} \in \mathbb{R}^{d \times N}$, the problem of sparse representation is to find a linear combination of a small number of entries or atoms in the dictionary to represent a frame, i.e.,

$$\hat{\mathbf{z}}_j^i = \arg \min \|\mathbf{z}_j^i\|_0 \quad s.t. \quad \mathbf{V}\mathbf{z}_j^i = \mathbf{x}_j^i, \quad (1)$$

where $\|\mathbf{z}_j^i\|_0$ is the ℓ_0 norm that equals to the number of nonzero entries in the vector $\mathbf{z}_j^i$. Finding the solution of Equation (1) is a NP hard problem. Fortunately, recent developments in sparse representation [11] reveal that if the solution is sparse enough, the solution of Equation (1) is equivalent to the solution by replacing the ℓ_0 norm with ℓ_1 norm, i.e.,

$$\hat{\mathbf{z}}_j^i = \arg \min \|\mathbf{z}_j^i\|_1 \quad s.t. \quad \mathbf{V}\mathbf{z}_j^i = \mathbf{x}_j^i, \quad (2)$$

The solution of Equation (2) can be efficiently estimated by optimizing

$$\hat{\mathbf{z}}_j^i = \arg \min \|\mathbf{x}_j^i - \mathbf{V}\mathbf{z}_j^i\|_2^2 + \lambda \|\mathbf{z}_j^i\|_1, \quad (3)$$

where λ is a sparsity regularization parameter.

There are two issues to be solved in the sparse coding scheme. The first one is how to construct the dictionary $\mathbf{V}$ and the other one is how to compute the sparse representation $\hat{\mathbf{z}}_j^i$ with a fixed dictionary. In [7], all the training images are concatenated to construct the dictionary. Similarly, we concatenate all the frames in the training videos as the dictionary, i.e., $\mathbf{V} = [\mathbf{x}_1^1, \mathbf{x}_1^2, \dots, \mathbf{x}_1^{T_1}, \dots, \mathbf{x}_M^1, \mathbf{x}_M^2, \dots, \mathbf{x}_M^{T_M}] \in \mathbb{R}^{d \times N}$. Note that in order to ensure the over-complete of the dictionary, the number of basis atoms N in the dictionary must be much larger than the dimension of atoms vectors d . This is always satisfied as the number of frames (i.e., $N = \sum_n^M T_j$) in training videos is much larger than the dimension of the pyramid HoG descriptor. Given the dictionary $\mathbf{V}$, the calculation of $\mathbf{z}_j^i$ in Equation (3) is a convex optimization problem. Many methods have been proposed to find the solution. In this work, we used the Orthogonal Matching Pursuit (OMP) method provided in [12].

3.2 Classification based on reconstruction errors

The SRC method in [7] classifies a new test sample to a class that corresponds to the minimum of reconstruct error. In this work, our target is to classify a new

video that contains a series of image frames. This can be achieved by extending the classification criterion by two ways. On one hand, each frame in a new test video is classified to an behaviour class based on the reconstruct error. The video is assigned to the class corresponding to the largest number of frames in the video. On the other hand, we alternatively sum the reconstruct errors of frames in the video as the overall residual, based on which the decision is made.

Similar to [7], for each class k , let δ_k be a function that selects all the coefficients corresponding to the k th class. For $\mathbf{z}_j^i \in \mathbb{R}^d$, $\delta_k(\mathbf{z}_j^i) \in \mathbb{R}^d$ is a vector whose nonzero entries are coefficients in $\mathbf{z}_j^i$ that associate with the k th class. The reconstruction error for a frame i with respect to the k th class is:

$$R_k(\mathbf{x}_j^i) = \|\mathbf{x}_j^i - \mathbf{V}\delta_k(\hat{\mathbf{z}}_j^i)\|_2. \quad (4)$$

Hence, a frame i is assigned to the behaviour class:

$$l(\mathbf{x}_j^i) = \arg \min_k R_k(\mathbf{x}_j^i). \quad (5)$$

Assume that the number of class is C , the number of frames in a video $\mathbf{X}_j$ that belongs to the class k is:

$$n_k(\sum_k n_k = T_j). \quad (6)$$

The video is classified to the class corresponding to the maximum number of frames:

$$c = \arg \max_k n_k. \quad (7)$$

We refer this classification method to as SC-A.

The SC-A makes the decision based on the class label of individual frames. Alternatively, we can determine the class of video based on the overall reconstruct error, which is defined as the sum of reconstruct errors of individual frames in the video, i.e.,

$$R_k(\mathbf{X}_j) = \sum_i^{T_j} R_k(\mathbf{x}_j^i). \quad (8)$$

The class of the video is determined as:

$$c = \arg \min_k R_k(\mathbf{X}_j). \quad (9)$$

This criterion can be regarded as a “soft” version of SC-A, since it sums all the reconstruct errors of individual frames as the overall reconstruct error. We refer this method to as SC-B. In the experimental section, we will compare the performance of the two methods.

4 Experiment

We evaluated the proposed method on the publicly available Weizmann dataset [13]. There are 10 behaviours (bend, jack, jump, pjump, run, side, skip, walk, one-hand wave, and two-hands wave) performed by 9 persons in the dataset. The same behaviour performed by different subjects has substantial variations. Fig. 1 shows example frames from the behaviour dataset. We directly used the tracked interest regions provided by

Figure 1. Example frames from the Weizmann dataset. The behaviours are: bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2 [13].

Figure 2. Confusion matrices for Weizmann dataset by SC-A (left) and SC-B (right).

the datasets, since human tracking and localization is not our main concern in this work.

We performed the leave-one-out evaluation on the dataset. All videos from the same subject are used for testing, while the others are treated as training data. This procedure is repeated for each of the subjects. The average recognition accuracy is reported. We divided the interest region into $l = 3$ levels and empirically set the number of gradient orientation bins as 8, giving a representation with a dimension of $d = 8 \sum_{l=0}^3 4^l = 680$. We employed the PCA to project it to a lower space of 200 dimensions. The sparsity regularization parameter was set to $\lambda = 1.2/\sqrt{\hat{d}}$, where $\hat{d}$ is the projected feature dimension (i.e., 200). The SC-A method obtains an average accuracy of 96.67%, while the SC-B achieves 97.78% on the dataset. The experimental results show that the SC-B method slightly outperforms the SC-A method. This is probably because that the SC-B utilizes the accumulated reconstruction errors of all the frames in a video as the overall error, while the SC-A makes the decision based on individual frame.

4.1 Confusion matrix

It is interesting to see which kinds of behaviours are more discriminative. Fig. 2 gives the confusion matrices by the SC-A and SC-B methods. Each row of the confusion matrix gives the probability that a certain behaviour is identified as a member of behaviour classes labelled by the columns. The diagonal elements in the matrix represent the correct classification rate of a certain behaviour. It can be seen that quite a small number of videos are misclassified, particularly, the behaviours “skip” and “run” are less discriminative than the others.

4.2 Comparison with the CRF and the HMM

Human motion is a temporal dynamic information. State-space models such as the Hidden Markov Model

Table 1. Activity recognition results

Models	Accuracy (%)
Hidden Markov Model	92.22
CRF($\omega = 0$)	94.44
CRF($\omega = 1$)	94.44
SC-A	96.67
SC-B	97.78

Table 2. Comparative results on Weizmann dataset

Methods	Accuracy (%)
Dollár et al. [2]	85.20
Niebles et al. [3]	90.00
Bregonzio et al. [5]	96.66
Ali and Shah [16]	95.75
Wang and Mori [4]	100
SC-A	96.66
SC-B	97.78

(HMM) [14] and the Conditional Random Filed (CRF) [15] are effective to capture this temporal order information. Here, we compared our method with the HMM and the CRF using the same pyramid HoG descriptor. The number of states of the HMMs is varied between 3 and 6. Only the highest recognition accuracy is reported for comparison. A linear-chain CRF with one step (i.e., $\omega = 0$) and long dependence (i.e., $\omega = 1$) are trained for classification.

Table 1 summarizes the recognition accuracy using the examined methods. It can be seen that the sparse coding method (both SC-A and SC-B) achieves highest accuracy, which shows that the proposed method performs better than HMM and CRF. It can be also observed that the CRF slightly outperforms the HMM in the experiment. However, the long-dependence of CRF does not improve the recognition accuracy. This is probably due to the over-fitting problem of the training data.

4.3 Comparison with other methods

We also compared our method with several existing state-of-the-art methods. All the methods are validated using the same leave-one-out validation protocol. The best results achieved by the methods are summarized in Table 2. It can be seen that our method performs very well relatively to the examined methods. It should be noticed that the dataset used in the method in [4] is an old version and only contains nine behaviours, while the new dataset has ten behaviours.

4.4 Robustness Test

In order to test the robustness of the approach, we used the Weizmann dataset as training data and tested the method on an additional dataset provided in [13]. The dataset contains a “walking” behaviour in ten difficult scenarios shown in Fig. 3. Three of the ten “walk-

Figure 3. Example frames of “walking” behaviour in different real scenarios. From top left to bottom right, the scene is : with a bag, with a briefcase, with a dog, with keens up, jump walk, moon walk, foot occlusion, normal walk occlusion by pole and in a skirt, respectively.

ing” sequences in the different scenarios dataset are misclassified, demonstrating that the sparse coding method is robust to different scenarios to some degree, although the test datasets are relatively simple and small.

5 Conclusion

In this work, we extended the sparse representation classification method in [7] to recognize human behaviour in video sequences. Each frame of a video is represented as a linear combination of a few elements in the dictionary. The class of a video is determined based on the reconstruction errors of all the frames in the video. Experimental results reveal that the proposed method achieves higher accuracy than those of most state-of-the-art methods.

REFERENCES

- [1] Ronald Poppe, “A survey on vision-based human action recognition,” *Image and Vision Computing*, vol. 28, pp. 976–990, 2010.
- [2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *VS-PETS*, 2005, pp. 65–72.
- [3] J. Niebles, H. Wang, and F-F. Li, “Unsupervised learning of human action categories using spatial-temporal words,” *Internatinal Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.
- [4] Y. Wang and G. Mori, “Hidden part models for human action recognition: Probabilistic versus max margin,” *IEEE Tran. PAMI*, vol. 33, no. 7, pp. 1310–1323, 2011.
- [5] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time interest points,” in *CVPR*, 2009, pp. 1948–1955.
- [6] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “On-line dictionary learning for sparse coding,” in *ICML*, 2009, pp. 689–696.
- [7] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE PAMI*, vol. 31, pp. 210–227, 2009.
- [8] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *CVPR*. Ieee, 2009, pp. 1794–1801.
- [9] Y. Fu Y. Liu Y. Zhu, X. Zhao, “Sparse coding on local spatial-temporal volumes for human action recognition,” in *ACCV*, 2010.
- [10] Yang Bai, Lihua Guo, Lianwen JIN, and Qinghua Huang, “A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition,” in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 3305–3308.
- [11] E.J. Candes and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *IEEE Trans. Inform Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.
- [12] Joel A. Tropp, Anna, and C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Inform. Theory*, vol. 53, pp. 4655–4666, 2007.
- [13] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” *IEEE Trans. PAMI*, vol. 29, no. 12, pp. 2247–2253, dec. 2007.
- [14] L.R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [15] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *ICML*, 2001, pp. 282–289.
- [16] S. Ali and M. Shah, “Human action recognition in videos using kinematic features and multiple instance learning,” *IEEE Tran. PAMI*, vol. 32, no. 2, pp. 288–303, 2010.