SPARSE CODING-BASED SPATIOTEMPORAL SALIENCY FOR ACTION RECOGNITION

Tao Zhang, Long Xu, Jie Yang and Pengfei Shi

Institute of Image Processing and Pattern Recognition
Shanghai Jiaotong University
Shanghai 200240, China

Wenjing Jia

Faculty of Engineering and Information Technology
University of Technology, Sydney
PO Box 123, Sydney, Australia

ABSTRACT

In this paper, we address the problem of human action recognition by representing image sequences as a sparse collection of patch-level spatiotemporal events that are salient in both space and time domain. Our method uses a multi-scale volumetric representation of video and adaptively selects an optimal space-time scale under which the saliency of a patch is most significant. The input image sequences are first partitioned into non-overlapping patches. Then, each patch is represented by a vector of coefficients that can linearly reconstruct the patch from a learned dictionary of basis patches. We propose to measure the spatiotemporal saliency of patches using Shannon's self-information entropy, where a patch's saliency is determined by information variation in the contents of the patch's spatiotemporal neighborhood. Experimental results on two benchmark datasets demonstrate the effectiveness of our proposed method.

Index Terms—Sparse coding, spatiotemporal saliency, action recognition, Shannon information entropy

1. INTRODUCTION

Recognizing human actions is a significant research area since a large amount of the information in image sequences is carried in the human action. Moreover, action recognition technology is continuously evolving and it is important to devise techniques to incorporate these improvements into existing systems. However, problems arisen from complex background, changing illumination, large variations in human appearance, different postures and body sizes within the same class have made this task very challenging.

The goal of action recognition is to recognize common human actions in real life settings. Common applications that make use of action recognition are health-assistive smart homes and smart environments, such as the Activities of Daily Living (ADLs) system [2] monitoring the functional health of a smart home resident [1, 2, 3], et al..

During the last few years, action recognition has been a hot topic. To some extent the problem has become tractable by using computer vision techniques. The common approach is to perform feature extraction from input video data and then feed the extracted features into a trained classifier for classification. Generally, features used to depict action video can be roughly divided into two groups, i.e. global features [1, 3] and local features [4, 5]. Methods based on global features first localize the person through foreground extraction or tracking and then extract features from the localized region. Methods based on local features consider a space–time video volume as a collection of local parts, where each one consists of some distinctive motion patterns. Each part is represented by local descriptors, which are then quantized into a vocabulary composed of visual words [6].

Among different approaches used for action recognition, detecting visual saliency in a video is considered a good way to understand the contents of the video [7, 8, 9, 10]. This is because successfully detecting visual saliency can substantially reduce the computational complexity of the whole action recognition process. Image saliency has been well explored in computer vision area, where salient-based detectors are normally based on different measures related to cornerness, entropy, global texture or periodicity. Recently, graph-based saliency [11], saliency in frequency and spatial domain [12], machine learning-based saliency [13] and global contrast-based saliency [14] have also been proposed. Nevertheless, there remain spaces of improvement using saliency-based approaches for successful recognition of human actions.

Recent studies in computer vision have demonstrated that sparse coding is an effective tool for image representation for various applications such as image classification, face recognition, image denoising, as well as saliency detection [2,15,16].

Inspired by the success related to sparse coding and saliency modelling, we study the incorporation of sparse image features and visual saliency measure. Aiming to develop an effective algorithm for action recognition, we derive a sparse coding-based salient feature extraction method for video. The derived spatiotemporal saliency can be regarded as a volumetric representation of the video where the sparse features interact to measure the spatiotemporal saliency.

The approach proposed in this paper follows a two-module framework, as illustrated in Fig. 1. The training module selects representative training data and extracts sparse features of image patches. Spatiotemporal saliency of

the sparse features is then measured for each patch. These saliency measures are then used for action classification.

In the context of this framework, our main contributions in this paper are two-fold: 1) We propose a sparse representation of image sequences as a collection of spatiotemporal events that are localized at patch level; 2) We define the spatiotemporal saliency for each patch using Shannon's self-information measure, which can examine the entropy behavior in spatiotemporal cylinders.

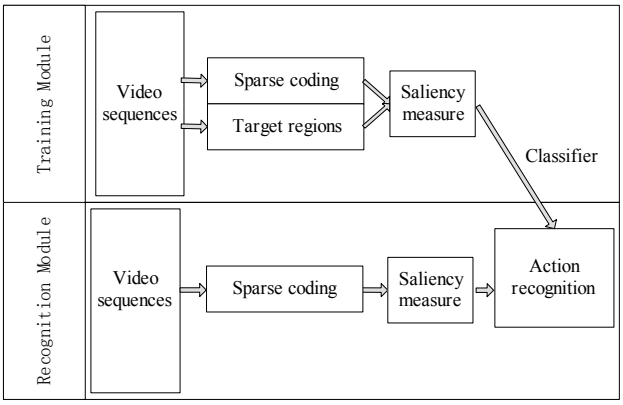

Fig. 1. The framework of the proposed method.

The rest of the paper is organized as follows. Sections 2 and 3 detail sparse feature representation and spatiotemporal saliency measurement respectively. Section 4 presents comparative experimental results with the state-of-the-art approaches. We conclude the paper in Section 5.

2. SPARSE FEATURE REPREENTATION

There exists much evidence indicating that the primate visual system is built on the principle of establishing a sparse representation of image statistics. If we treat sparse dictionary as a filter, the process of sparse coding is equivalent to image filtering, and the result of sparse coding can be regarded as the extracted features of the input image.

In this paper, sparse feature is determined by quantifying the self-information of each local image patch. Each image patch is projected into the space of a dictionary of image patches (basis functions) learned from a repository of activity scenes. Each patch is then represented by a vector of basis coefficients that can linearly reconstruct the patch.

Mathematically, given a set of training data $Y = [y_1, y_2, \cdots, y_q] \in R^{m \times q}$, the $n \times m$−dimensional sparse dictionary $D = [d_1, d_2, \cdots, d_n] \in R^{m \times n}$ can be found by solving an empirical cost function minimization problem as:

$$\min_{a \in R^n} \frac{1}{q} \sum_{j=1}^{q} \left(\frac{1}{2} \left\| y_j - D a_j \right\|_{l_2}^2 + \lambda \left\| a_j \right\|_{l_1} \right) \qquad (1)$$

where $\|\bullet\|_{l_1}$ and $\|\bullet\|_{l_2}$ denote l_1 and l_2-norm respectively, λ is a positive regularization parameter controlling the trade-off between fitting degree and sparseness, and a_j is the corresponding sparse representation of vector y_j.

We represent an image patch by a linear combination of some basis functions which act as feature detectors. Given an input image, it is first resized to $2^w \times 2^w$ pixels where the patch size *w* is selected in a way that 2^w is divisible to *w*. Then, using Eq. 1, the coefficients that linearly reconstruct each patch are calculated and used to represent that patch. By reshaping the reconstructed patches and aligning them, the original image can be reproduced.

The dictionary *D* contains atoms that represent the basic patterns of the specific data distribution in feature space. Given the dictionary $D = [d_1, d_2, \cdots, d_n] \in R^{m \times n}$, the sparse coding, denoted as a^*, of an input signal $x \in R^m$ can be found by solving an l_1 and l_2-norm minimization problem:

$$a^*(x, D) = \arg\min_{a \in R^n} \frac{1}{2} \left\| x_i - D a_j \right\|_{l_2}^2 + \lambda \left\| a_j \right\|_{l_1}, \qquad (2)$$

where $x \approx x^* = Da^*$ (x^* is the estimation of *x*). The optimization over a^* is convex when the dictionary *D* is constant. To seek a sparse a^*, the LARS-lasso approach [15, 16] is employed to solve Eq. 2.

Let $X = \{x_1, x_2, \cdots, x_n\}$ represent the set of linearized image patches with no overlapping. Using Eq. 2, we can obtain the sparse feature values of each patch of the image *X*, denoted as $FX = [a_1^*, a_2^*, \cdots, a_N^*] \in R^{n \times N}$.

To learn a dictionary of basis patches (i.e. minimizing Eq. 1), we extracted 100,000 8×8 image patches (for each channel of RGB) from randomly selected color images from action scenes. Each basis function in the dictionary is a vector of 8 × 8 = 64 dimension. We have experimented different dictionary sizes, i.e. 50, 100, 150, 300, 500, and 1000, and chose 500 considering accuracy and computation complexity. The sparse codes a_j are computed with the above basis using the LARS algorithm [15, 16].

3. MEASURING SPATIOTEMPORAL SALIENCY

Saliency computation in video is a problem of assigning a measure of interest to each spatiotemporal visual unit. Our saliency model is based on Shannon entropy. We calculate the Shannon entropy of each patch for each RGB channel and fuse the saliency maps of each channel to generate a saliency map.

In order to detect spatiotemporal salient patches at peaks of activity variation, we consider the cylindrical spatiotemporal neighborhoods of a patch at different spatial radii *s* and temporal depths *d*. More specifically, let us denote by $N_d(s^*, v^*)$ the set of patches in a cylindrical neighborhood of scale $s^* = (s, d)$ centered at the spatiotemporal patch v^*

= (*x*, *y*, *t*) in the given image sequence. At each patch v^* and for each scale s^*, we define the spatiotemporal saliency $y_D(s^*, v^*)$ by measuring the information changes in the contents within $N_d(s^*, v^*)$. Here, we apply the sparse feature *FX* (obtained in Section 2) in the temporal domain in order to achieve a measure of actions. The input signal that we use is the sparse feature *FX*.

For each patch v^* = (*x*, *y*, *t*) in the image sequence, let us calculate the Shannon entropy of the signal histogram in a spatiotemporal neighborhood around the patch. The signal's Shannon entropy $H_D(s^*, v^*)$ in the spatiotemporal neighborhood $N_d(s^*, v^*)$ is given by:

$$H_D(s^*, v^*) = -\int_{q \in FX} PD(s*, v*) \log_2 PD(s*, v*) dq \,, \qquad (3)$$

where $PD(s*, v*)$ is the probability density of the signal as a function of scale s^* and position v^*, and q denotes the signal value obtained from Eq. 2. Here, we use histogram to approximate the probability density $PD(s*, v*)$.

We adopt the automatic selection method in [8] to determine the optimal scale. More specifically, we consider the scales at which the entropy value reaches local maxima as candidate salient scales. Let us denote the set of scales at which the entropy is peaked as ${S_p}^*$, which is defined as:

$$\begin{aligned} {S_p}^* = \Bigg\{ s^* : \frac{\partial H_D(s^*, v^*)}{\partial s} = 0 \wedge \frac{\partial^2 H_D(s^*, v^*)}{\partial s^2} < 0 \wedge \\ \frac{\partial H_D(s^*, v^*)}{\partial d} = 0 \wedge \frac{\partial^2 H_D(s^*, v^*)}{\partial d^2} < 0 \Bigg\} \end{aligned} \qquad (4)$$

where "^" denotes logical anding operation.

Then, we define the saliency metric as:

$$y_D(s^*, v^*) = H_D(s^*, v^*) W(s^*, v^*), \forall s^* \in {S_p}^* . \qquad (5)$$

Here, the first term measures the variation in the sparse information content of the signal. The weighting function $W(s^*, v^*)$ measures how prominent the local maxima is at s^*, and is defined by

$$\begin{aligned} W(s^*, v^*) = s \int_{q \in FX} \left| \frac{\partial PD(s*, v*)}{\partial s} \right| dq + \\ d \int_{q \in FX} \left| \frac{\partial PD(s*, v*)}{\partial d} \right| dq, \forall s(s, d) \in {S_p}^* \end{aligned} \qquad (6)$$

When a peak in the entropy for a specific scale is distinct, the probability density functions of the corresponding patches at neighboring scales will differ substantially, giving a big value to the integrals of Eq. 6 and thus to the corresponding weight value assigned. On the contrary, when the peak is smoother, the integrals in Eq. 6 and the corresponding weight will have a smaller value.

Eq. 5 gives an overall measure of how salient a spatiotemporal patch v^* is at certain candidate scale s^*. Using the saliency feature-detection scheme described above, we represent a given image sequence by a set of features, where each feature corresponds to a cylindrical salient region of the image sequence in the space–time domain. The final result is a hierarchy of video volumes that represent the input sequence in decreasing spatiotemporal scales. These are then used to train a classifier for action recognition.

4. EXPERIMENTAL RESULTS

We evaluate the proposed model by setting up experiments using two public datasets widely used in the action recognition domain, i.e. the KTH dataset [2] and the Weizmann dataset [2]. We first briefly introduce the two datasets and then devote the rest of the section to quantitative analysis of the proposed method in comparison with existing methods. The regularization parameter λ in Eq. (1) and Eq. (2) is set to $\frac{1.2}{\sqrt{m}}$ according to [16], where *m* is the dimension of the original feature.

4.1. Datasets

The Weizmann Dataset [2]: This dataset contains 93 low-resolution video clips (180×144 pixels) from nine different subjects, each of whom performs 10 different actions including walking (walk), running (run), jumping (jump), galloping sideways (side), bending (bend), one-hand-waving (waveone), two-hands-waving (wavetwo), jumping in place (pjump), jumping jack (jack), and skipping twice. The camera setting is fixed with no occlusion or viewpoint change. Each subject performs under similar plain background.

The KTH Dataset [2]: This dataset is relatively complex and can be considered as an important benchmark dataset to evaluate various human action recognition algorithms. It contains six actions: walking (walk), jogging (jog), running (run), boxing (box), hand waving (hwav) and hand clapping (hclap), performed by twenty-five subjects in four different scenarios including indoor, outdoor, changes in clothing and variations in scale. It contains 599 low-resolution video clips (160×120 pixels), for one of the videos is missing.

4.2. Recognition results on the Weizmann dataset

For fair comparison, the same framework of [5] and [7], which is based on bag-of-words (BoW) and a Nearest-Neighbor Classifier (NNC), has been used. The dataset is divided into a training set and a testing set. We create a codebook $W = \{w_k\}, k = 1, \cdots, K$ of K visual words using *k*-means clustering on the set of descriptors from all patches. The salient patches of a given clip are associated to the most

similar visual word and a histogram h_m of visual word occurrence is extracted for the clip. We compute a leave-one-out estimate of the error by training on all persons minus one, testing on the remaining one and repeating the procedure for all persons.

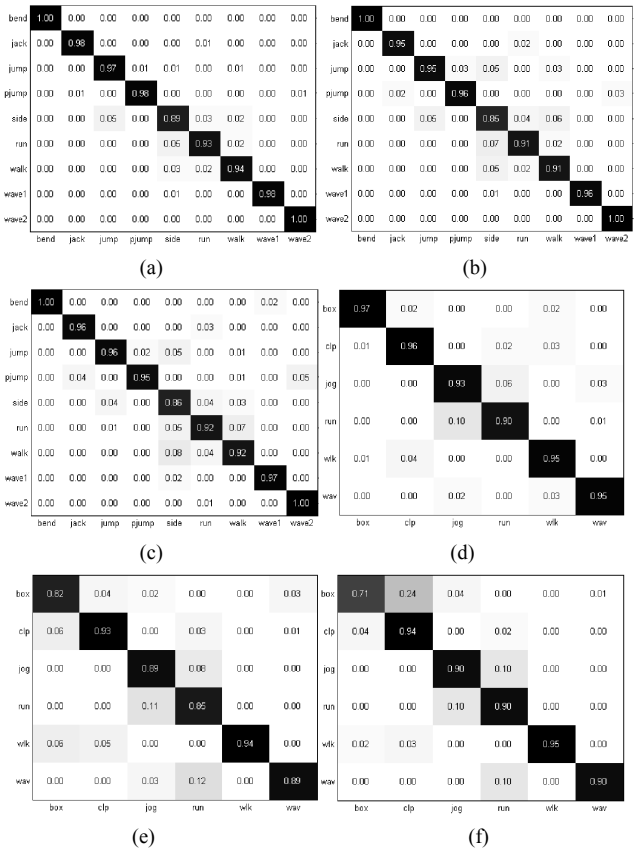

Fig. 2. Confusion matrices of results on the Weizmann dataset (a, b, c) and the KTH dataset (d, e, f). (a) Our model (overall: 96.3%), (b) Chen et al. [5] (overall: 94.3%), (c) Rapantzikos et al. [7] (overall: 94.9%), (d) our model (overall: 94.3%), (e) Chen et al. [5] (overall: 88.6%) and (f) Rapantzikos et al. [7] (overall: 88.3%).

Fig. 2 (a) shows the confusion matrix of our recognition results in comparison with those of two other methods ((b) and (c)) obtained on the Weizmann dataset. It can be seen that, our detector performs better than the other two over all actions with an overall rate of 96.3%. The recognition rate for some actions are high up to 100%, such as ‘‘bend’’ and ‘‘wave2’’.

4.3. Recognition results on the KTH dataset

Similar to Section 4.2, we follow the same framework and all evaluations were done with 5-fold cross-validation: four folds are used for training and one for testing. Fig. 2(d) shows the confusion matrix of our recognition results in comparison with the methods of [5] (Fig. 2(e)) and [7] (Fig. 2(f)) obtained on the KTH dataset.

Compared with the Weizmann dataset, the KTH dataset is more complex, so the recognition rate is lower than that on the Weizmann dataset. As is shown, four action classes are perfectly recognized. The other two action classes, such as “jog” and “run”, are relatively difficult to be recognized. Furthermore, a rather small confusion occurs between “wave” and “clap”, mainly because both behaviors involve the motion of hands. Overall, our detector performs better than the other two with an overall recognition rate of 94.3%. It achieves rates equal to or higher than 93% for all actions except for “run”. It seems that the inherent periodicity of these actions is well represented by our sparse coding-based salient feature extraction method.

Table 1 summarizes the results on the KTH dataset published in recent years. Our detector using NNC has achieved the best results among all these methods in comparison.

By verifying the obtained results, we can find that our proposed system is effective and robust for correct recognition of human actions.

Table 1. Performance comparison on the KTH dataset.

Algorithms	Accuracy	Classifier
Weinland et al. [3]	91.29%	SVM
Chen et al. [5]	88.61%	NNC
Laptev et al. [6]	91.80**%**	mc-SVM
Rapantzikos et al. [7]	88.34**%**	NNC
Ours	**94.33%**	NNC

5. CONCLUSION

In this paper, we extended the concept of saliency to the spatiotemporal domain in order to represent human motion with a sparse set of spatiotemporal features. We propose a sparse saliency representation of image sequences as a collection of spatiotemporal events. Different from traditional saliency-based approaches, our constructed sparse feature aims to detect saliency of patches rather than raw pixel values. We propose to measure spatiotemporal saliency of patches using Shannon's self-information entropy. Experimental results have shown a high recognition rate, and the choice of the sparse spatiotemporal patches saliency has a profound impact on the performance of recognition.

6. ACKNOWLEDGMENT

This research is partly supported by NSFC, China (No: 61273258) and 973 Plan，China (No. 2015CB856004).

7. REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," Image Vision Comput, vol. 28, no. 6, pp. 976–990, 2010.
[2] J.K. Aggarwal and M.S. Ryoo, "Human Activity Analysis: A Review," ACM Comput Surv, Volume. 43, pp. 1-43, 2011.
[3] D. Weinland and E. Boyer and L. Rhône-alpes, "Action recognition using exemplar-based embedding," CVPR, pp. 1–7, 2008.

[4] I. Laptev and T. Lindeberg, "Space-time interest points," ICCV, pp. 432–439, 2003.
[5] M. Y. Chen and A. Hauptmann, "MoSIFT: Recognizing Human Actions in Surveillance Videos," Tech. rep., Carnegie Mellon University, 2009.
[6] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," CVPR. pp. 1–8, 2008.
[7] K. Rapantzikos, Y. Avrithis, S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," CVPR, pp.1-8, 2009.
[8] A. Oikonomopoulos, I. Patras and M. Pantic, "Spatiotemporal Salient Points for Visual Recognition of Human Actions," IEEE Trans. Systems, Man, and Cybernetics, Part B, Vol. 36, No. 3, pp. 710-719, Jun. 2006.
[9] D. Rudoy, D.B Goldman, et al., "Learning video saliency from human gaze using candidate selection", CVPR, pp. 4321-4328, 2013.
[10] N. D. Bruce and J. K. Tsotsos, " Saliency, attention, and visual search: An information theoretic approach," Journal of Vision, Vol. 36, No. 3, pp. 1-24, 2009.
[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," NIPS, pp. 545-552, 2007.
[12] J. Li, M. D. Levine, X. An, et al., "Saliency detection based on frequency and spatial domain analysis," BMVC, Dundee, vol. 86, pp. 1-11, 2011.
[13] T. Liu, Z. Yuan, J. Sun, et al., "Learning to detect a salient object," PAMI, vol. 33, no. 2, pp. 353-367, 2011.
[14] M. M. Cheng, G. X. Zhang, N. J. Mitra, et al. , "Global contrast based salient region detection," CVPR, pp. 409-416, 2011.
[15]J. C. Yang, C. Yu, H. Thomas: Supervised translation-invariant sparse coding. CVPR , pp. 3517-3524, 2011.
[16] G. Mairal, F. Bach, J. Ponce, G. Sapiro: Online dictionary learning for sparse coding. ICML, pp. 689-696, 2009.