Activity Recognition in Still Images with Transductive Non-negative Matrix Factorization

Naiyang Guan[1](✉), Dacheng Tao[2], Long Lan[1], Zhigang Luo[1],
and Xuejun Yang[3]

[1] Science and Technology on Parallel and Distributed Processing Laboratory,
College of Computer, National University of Defense Technology,
Changsha, People's Republic of China
{ny_guan,zgluo}@nudt.edu.cn, lan19901@126.com

[2] Centre for Quantum Computation and Intelligent Systems and the Faculty
of Engineering and Information Technology, University of Technology,
Sydney, Australia
dacheng.tao@uts.edu.au

[3] State Key Laboratory of High Performance Computing,
National University of Defense Technology, Changsha, People's Republic of China
xjyang@nudt.edu.cn

Abstract. Still image based activity recognition is a challenging problem due to changes in appearance of persons, articulation in poses, cluttered backgrounds, and absence of temporal features. In this paper, we proposed a novel method to recognize activities from still images based on transductive non-negative matrix factorization (TNMF). TNMF clusters the visual descriptors of each human action in the training images into fixed number of groups meanwhile learns to represent the visual descriptor of test image on the concatenated bases. Since TNMF learns these bases on both training images and test image simultaneously, it learns a more discriminative representation than standard NMF based methods. We developed a multiplicative update rule to solve TNMF and proved its convergence. Experimental results on both laboratory and real-world datasets demonstrate that TNMF consistently outperforms NMF.

Keywords: Still image based action recognition · Non-negative matrix factorization · Transductive learning

1 Introduction

Activity recognition aims to recognize actions and goals of one or more individuals from a series of observations on the individuals' actions and the environmental conditions. It has found many applications in human-computer interaction [29], human interaction recognition [28], robot trajectory planning [30], and video surveillance [31] thanks to the convenience of capturing videos through cameras [1–3,13,27]. Until now, activity recognition is an open and challenging problem due to changes in appearance of persons, articulation in poses, cluttered backgrounds, and camera movements.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 802–817, 2015.
DOI: 10.1007/978-3-319-16178-5_56

Recognizing actions from benchmark videos has achieved promising performance because of the dynamic features, but it is difficult to recognize actions recorded in still wild images, e.g., images collected from Internet, because the dynamic features cannot be extracted from still images. To recognize actions from still images, it is important to extract representative cues including both high-level and low-level cues. Traditional video-based activity recognition can directly use the low-level cues such as the spatiotemporal interest point [23] extracted from space-time volume, but the still image-based activity recognition usually cannot because only the spatial information is available on single images [35]. The high-level cues can be characterized by various low-level features, e.g., color names [21], and different high-level cues can be combined to enhance the performance, e.g., combining pose and context information [20]. Interested readers can refer to [24] for a systematic survey.

To construct high-level cues, it is an important pre-processing step to detect human bodies, body parts and objects. However, it is quite challenging because existing object detection methods usually work unsatisfactorily. Liu *et al.* [11] proposed to represent actions by selecting key poses from video sequences. Zhang and Tao [12] proposed the slow-feature analysis (SFA) framework to recognizing human actions from video sequences by incorporating discriminative information with SFA and spatial relationship of body components. Although these methods have achieved great successes by utilizing human poses, they are not direct for action recognition in still images due to the difficulty to extract body components. In this paper, we constructed a high-level cues by clustering human poses with non-negative matrix factorization (NMF, [7]) to avoid explicitly reasoning about the body components [22]. Non-negative matrix factorization is a popular data representation method which can extract intrinsic structure of dataset and boost the performance of subsequent processing. Different from conventional data representation methods, e.g., principal component analysis (PCA, [14]) and Fisher's linear discriminative analysis (FLDA, [15]), which learns holistic representation, NMF can learn parts-based representations from non-negative datasets. For example, it can extract several versions of facial components such as 'noses', 'eyes', and 'mouth' from frontal face image datasets. It is therefore reasonable to believe that NMF can automatically extract body poses from bounding boxes.

Thurau and Hlavac [4] proposed static histogram of oriented gradient (HOG)-based features for activity recognition on still images by clustering a set of training human poses with NMF and utilizing histograms of the clustered poses to represent each action. At the classification stage, they concatenated the pose clusters of all actions and features of background, and calculated the histogram of each test image on concatenated features and determined the label by classification. Since then, many works utilize NMF in activity recognition. Agarwal and Xia [5] applied NMF to 3D poses recovery problem since NMF can effectively represents local features of human body. According to [5], background usually has a negative influence on action recovery because its changes are usually misunderstood as human actions. NMF is suitable for recovering poses from single image because it can significantly separate background from action poses. Waltner *et al.* [6] utilized NMF to recognize actions from a small amount of video frames. Different from [4] and [5], their method considers HOG of both appearances and motions. The discriminative power of the learned poses is improved by motions, but it is far from enough because aforementioned methods [4–6] ignore test samples during training.

In this paper, we propose a novel method to recognize actions from still images by using transductive NMF (TNMF). TNMF jointly learns a dictionary of features on both training images from different actions and the test image to be recognized. In particular, TNMF has two types of objectives: 1) it minimizes the distance between the visual descriptors of the training poses of each action and the product of its features and encodings, and 2) it minimizes the distance between the visual descriptor of test image and the product of dictionary concatenated by those features of all actions and an encoding vector. Intuitively, since the dictionary of features learned by TNMF contains the visual features from both training images and test image, it can more accurately recover the pose in single still image, and thus boost the recognition performance. TNMF balances both objectives by a positive parameter and utilizes a multiplicative update rule (MUR) to learn all features and the corresponding encodings. In this paper, we proved the convergence of the MUR-based algorithm for TNMF. Experiment results on both laboratory datasets and real-life datasets confirm that TNMF significantly outperforms NMF in still image-based activity recognition.

This paper is organized as follows: Section 2 surveys both NMF and its application in activity recognition; we introduce the TNMF model and its MUR based algorithm in Section 3; Section 4 verifies the method on both laboratory and real-world datasets and Section 5 concludes this paper.

2 Related Works

2.1 Non-negative Matrix Factorization

Given a non-negative dataset, i.e., $V \in \mathbb{R}_{+}^{m\times n}$, non-negative matrix factorization (NMF, [7]) decomposes it into the product of two lower-rank matrices, i.e., $W \in \mathbb{R}_{+}^{m\times r}$ and $H \in \mathbb{R}_{+}^{r\times n}$, where $r \ll \min\{m, n\}$, by solving the following problem

$$\min_{W\geq 0, H\geq 0} ||V - WH||_F^2. \tag{1}$$

Usually, W and H can be considered as features and encodings, respectively. It is obvious that NMF represents each sample by only additive, non-subtractive combination of features. Therefore, NMF yields parts-based features representation.

Since such parts-based representation has strong evidence in human brain, NMF has been widely applied in many real-world applications such as text mining [8–10,14] and hyper-spectral imaging [10,15].

2.2 Transductive NMF

Recently, Guan *et al.* [16] have proposed transductive NMF (TNMF) to simultaneously learn from multiple tasks, i.e., V_k, where $1 \leq k \leq K$. TNMF combines both training stage and test stage together to simultaneously learn single features for each task and coefficient of test sample on concatenated dictionary.

The objective function of TNMF is

$$\min_{\forall 1 \leq k \leq K, W_k \geq 0, H_k \geq 0, \overline{H} \geq 0} \{\sum_{k=1}^{K} ||V_k - W_k H_k||_F^2 + \lambda ||\overline{V} - \overline{WH}||_F^2\}, \tag{2}$$

where $\overline{W} = [W_1, \cdots, W_K]$, and $\lambda \in [0, 1]$ is a positive tradeoff parameter. When $\lambda = 0$, TNMF reduces to NMF on each task separately.

2.3 NMF-Based Activity Recognition

Taking the advantage of clustering ability the parts-based representation of NMF, Thurau and Hlavac [4,6] proposed a static HOG-based NMF method for activity recognition on still images since the HOG-descriptor of any an image is non-negative. Given training HOG-descriptors of all actions, i.e., V_k for the k-th action of totally K actions, they utilized NMF to learn features W_k and encodings H_k, i.e., by

$$\min_{W_k \geq 0, H_k \geq 0} ||V_k - W_k H_k||_F^2. \tag{3}$$

By concatenating features of all actions together, they constructed a dictionary of features, i.e., $\overline{W} = [W_1, \cdots, W_K]$, and projected the HOG-descriptors of test image, i.e., $\overline{V}$, onto $\overline{W}$ to calculate its encoding, i.e.,by

$$\overline{H} = \arg\min_{H \geq 0} ||\overline{V} - \overline{W}H||_F^2, \tag{4}$$

where $\overline{H}$ is the encodings of $\overline{V}$.

At the classification stage, they calculated the histogram of each action based on $\{H_1, \cdots, H_K\}$, and the histogram of the test image based on $\overline{H}$, followed by classification with the nearest neighbor (NN) classifier. Since the training stage of learning the features of each action (see the formula (3)) and the classification stage of learning the encodings on the dictionary of concatenated features (see the formula (4)) are separate, NMF usually suffers from overfitting problem.

3 TNMF-Based Activity Recognition in Still Images

In still image-based activity recognition, most actions have sufficient training images but some actions has rare training images because the training images are collected from different sources and the activities are performed separately by different individuals. Therefore, NMF cannot accurately learn features on limited training images due to the overfitting problem in this situation.

Since TNMF leverages the test set to enhance representing the training samples, it learns more representative dictionary and reduces the influence of overfitting by jointly learning the dictionary from both training and test sets. In other words, TNMF has better generalization ability than NMF. In this paper, we taken this advantage of TNMF to solve the overfitting problem in still image-based activity recognition [4]. In particular, we applied TNMF to jointly learns a dictionary on both training samples V_k, e.g., HOG-descriptors, from different actions and the test samples $\overline{V}$, e.g., HOG-descriptors, of the probe image to be recognized. Since TNMF transduces the training poses to the learned dictionary by incorporating the second term in (4), it represents the test data more accurately and overcomes the deficiency of NMF. Experimental results confirm that TNMF greatly boosts the activity recognition performance.

Although the objective function of TNMF is jointly non-convex with respect to all variables, i.e., $\{W_1, \cdots, W_K, H_1, \cdots, H_K, \overline{H}\}$, it is convex with respect to each of them separately. According to [17,20], we utilized the majorization minimization (MM) method to derive a multiplicative update rule (MUR, [16]) for solving TNMF (2). MUR updates W_k, H_k, and $\overline{H}$, respectively, by

$$W_k \leftarrow W_k \circ \frac{(V_k H_k^T + \lambda \overline{V} \overline{H}_k^T)}{(W_k H_k H_k^T + \lambda \overline{W} \overline{H} \overline{H}_k^T)}, \tag{5}$$

$$H_k \leftarrow H_k \circ \frac{W_k^T V_k}{W_k^T W_k H_k}, \tag{6}$$

and

$$\overline{H} \leftarrow \overline{H} \circ \frac{\overline{W}^T \overline{V}}{\overline{W}^T \overline{W H}}, \tag{7}$$

where $\circ$ signifies the element-wise multiplication operator, and $\overline{H}_k$ is the k-th component of $\overline{H}$ that corresponds to W_k, i.e., $\overline{H} = [\overline{H}_1^T, \cdots, \overline{H}_K^T]^T$. MUR alternatively updates all variables until they do not change the objective value of (2).

Distinguished from our previous work, we proved the convergence of the MURs (5)(6)(7) by using the majorization minimization (MM, [17]) technique. MM builds an auxiliary function whose curve lies above that of the original objective function everywhere and both curves are tangent at a certain point. When calculating the gradient of the original function is non-trival, MM instead updates the current variable by using the minimum of the constructed auxiliary function. The auxiliary function is defined in **Definition 1** and has the property shown **Lemma 1**.

Definition 1. Given x^t, the function $g(x, x^t)$ is an auxiliary function of $f(x)$, if $g(x, x^t) \geq f(x)$ and $g(x^t, x^t) = f(x^t)$.

Lemma 1. If $g(x, x^t)$ is an auxiliary function of $f(x)$, then $f(x)$ is non-increasing under the update rule $x^{(t+1)} = \arg\min_x g(x, x^t)$.

Proof. $f(x^{t+1}) \leq g(x^{t+1}, x^t) \leq g(x^t, x^t) = f(x^t)$. ■

It is easy to verify that (6) and (7) decrease the objective function because they are same as the MURs in [17]. It is remaining to prove that (5) decreases the objective function (see **Proposition 1**).

Proposition 1. The multiplicative update rule (5) decreases the objective function of (2).

Proof. At the t-th iteration round, we expect to prove that the update of W_k can decrease the objective function

$$f_t = \sum_{l \neq k}^{K} ||V_l - W_l^t H_l^t||_F^2 + ||V_k - W_k H_k^t||_F^2 + \lambda ||\overline{V} - \overline{W}^t \overline{H}^t + W_k \overline{H}_k^t - W_k \overline{H}_k^t||_F^2,$$

with all variables except W_k fixed. Since the first term does not influence f_t, it is only necessary to prove that (5) decreases the following objective function

$$f(W_k) = ||V_k - W_k H_k^t||_F^2 + \lambda ||\overline{V} - \overline{W}^t \overline{H}^t + W_k \overline{H}_k^t - W_k \overline{H}_k^t||_F^2. \tag{8}$$

To this end, we constructed its auxiliary function as follows:

$$\begin{aligned} g(W_k, W_k^t) &= f(W_k^t) + \langle \nabla f(W_k^t), W_k - W_k^t \rangle \\ &+ \langle \frac{W_k^t H_k^t {H_k^t}^T + \lambda \overline{W}^t \overline{H}^t {\overline{H}_k^t}^T}{(W_k^t)}), [W_k - W_k^t]^2 \rangle, \end{aligned} \tag{9}$$

where $\nabla f(W_k^t) = (W_k H_k^t - V_k) {H_k^t}^T + \lambda (\overline{W}^t \overline{H}^t - \overline{V}) {H_k^t}^T$ and $[\cdot]^2$ signifies the element-wise square of a matrix. Since it is obvious that $g(W_k^t, W_k^t) = f(W_k^t)$, we only need to show $f(W_k) \leq g(W_k, W_k^t)$ for any W_k.

To do this, we have the Taylor series expansion of $f(W_k)$ at W_k^t, and the objective function with respect to the (i, j)-th element of W_k is

$$\begin{aligned} f([W_k]_{ij}) &= f([W_k^t]_{ij}) + [\nabla f(W_k^t)]_{ij} ([W_k]_{ij} - [W_k^t]_{ij}) \\ &+ ([H_k^t {H_k^t}^T]_{jj} + \lambda [\overline{H}_k^t {\overline{H}_k^t}^T]_{jj}) ([W_k]_{ij} - [W_k^t]_{ij})^2. \end{aligned} \tag{10}$$

Since $H_k^t \leq 0$ and $W_k^t \leq 0$, we have

$$[H_k^t {H_k^t}^T]_{jj} \leq \frac{\sum_l [W_k^t]_{il} [H_k^t {H_k^t}^T]_{lj}}{[W_k^t]_{ij}} = \frac{[W_k^t H_k^t {H_k^t}^T]_{ij}}{[W_k^t]_{ij}}. \tag{11}$$

Since $\overline{H}_k^t \leq 0$ and $W_k^t \leq 0$, we have

$$[\overline{H}_k^t {\overline{H}_k^t}^T]_{jj} \leq \frac{\sum_l [W_k^t]_{il} [\overline{H}_k^t {\overline{H}_k^t}^T]_{lj}}{[W_k^t]_{ij}} = \frac{[W_k^t \overline{H}_k^t {\overline{H}_k^t}^T]_{ij}}{[W_k^t]_{ij}} \leq \frac{[\overline{W}_k^t \overline{H}_k^t {\overline{H}_k^t}^T]_{ij}}{[W_k^t]_{ij}}, \tag{12}$$

where the last inequality comes from the fact that $\bar{W}^t\bar{H}^t = \sum_{l\neq k}^K W_l^t\bar{H}_l^t + W_k^t\overline{H}_k^t$ and $\sum_{l\neq k}^K W_l^t\overline{H}_l^t \geq 0$.

By substituting (11) and (12) into (10), we can easily verify that $f(W_k) \leq g(W_k, W_k^t)$, and thus $g(W_k, W_k^t)$ is an auxiliary function of $f(W_k)$ according to

Definition 1. By setting $\frac{\partial g(W_k, W_k^t)}{\partial [W_k]_{ij})} = 0$ and substituting $\nabla f(W_k^t) = (W_kH_k^t - V_k){H_k^t}^T + \lambda(\overline{W}^t\overline{H}^t - \overline{V}){\overline{H}_k^t}^T$, we have

$$\begin{aligned} &[W_k^tH_k^t{H_k^t}^T]_{ij} - [V_k{H_k^t}^T]_{ij} + \lambda[\overline{W}^t\overline{H}^t{\overline{H}_k^t}^T]_{ij} - \lambda[\overline{V}{\overline{H}_k^t}^T]_{ij} \\ &\frac{[W_k^tH_k^t{H_k^t}^T]_{ij} + \lambda[\overline{W}^t\overline{H}^t{\overline{H}_k^t}^T]_{ij}}{[W_k^t]_{ij}}([W_k]_{ij} - [W_k^t]_{ij}) = 0. \end{aligned}$$

It is equivalent to

$$-[V_k{H_k^t}^T]_{ij} - \lambda[\overline{V}{\overline{H}_k^t}^T]_{ij} + \frac{[W_k^tH_k^t{H_k^t}^T]_{ij} + \lambda[\overline{W}^t\overline{H}^t{\overline{H}_k^t}^T]_{ij}}{[W_k^t]_{ij}}[W_k]_{ij} = 0. \quad (13)$$

From (14), we have the minimum of $g(W_k, W_k^t)$ with respect to the (i,j)-th element of W_k as follows:

$$[W_k^*]_{ij} = [W_k^t]_{ij}\frac{([V_k{H_k^t}^T]_{ij} + \lambda[\overline{V}{\overline{H}_k^t}^T]_{ij})}{[W_k^tH_k^t{H_k^t}^T]_{ij} + \lambda[\overline{W}^t\overline{H}^t{\overline{H}_k^t}^T]_{ij}}. \quad (14)$$

By rewriting (14) in a matrix form, we have

$$W_k^* = W_k^t \circ \frac{V_k{H_k^t}^T + \lambda\overline{V}{\overline{H}_k^t}^T}{W_k^tH_k^t{H_k^t}^T + \lambda\overline{W}^t\overline{H}^t{\overline{H}_k^t}^T}.$$

By setting $W_k^{t+1} = W_k^*$, we know that $f(W_k^{t+1}) \leq f(W_k^t)$ according to **Lemma 1**. This completes the proof. ■

Interestedly, the above proof procedure suggest the generalization ability of TNMF. By simple algebra, the formula (9) is equivalent to the following minimization:

$$\min_{W_k \geq 0} ||X_k - W_kY_k||_F^2,$$

where $X_k = [V_k, \sqrt{\lambda}(\overline{V} - \overline{W}^t\overline{H}^t + W_k^t\overline{H}_k^t)]$ and $Y_k = [H_k^t, \sqrt{\lambda}\overline{H}_k^t]$. It means that TNMF learns dictionary both from training examples and test examples. In other words, TNMF achieves better generalization ability than the standard NMF only on training examples.

Since MURs decrease the objective function of TNMF, the objective function gets more and more close to the minimum, and gets farther and farther from the initial point, on its fly. We therefore gave the following stopping condition of MUR like [25,26]:

$$\frac{|f_t - f_{t-1}|}{|f_t - f_0|} \leq \varepsilon, \tag{15}$$

where $f_t = \sum_{k=1}^{K} ||V_k - W_k^t H_k^t||_F^2 + \lambda ||\overline{V} - \overline{W}^t \overline{H}^t||_F^2$ signifies the objective value at the t-th iteration round ($t \leq 1$), and ε signifies the tolerance, i.e., $\varepsilon = 10^{-3}$. We summarized the total procedure of MUR for TNMF in **Algorithm 1**.

Algorithm 1. MUR for Optimizing TNMF

```
      Input: $\{V_1, \cdots, V_K\}$, $\overline{V}$, and $r$
      Output: $\{W_1, \cdots, W_K\}$, $\{H_1, \cdots, H_K\}$, and $\overline{H}$
1.    Initialize $\{W_1, \cdots, W_K\}$, $\{H_1, \cdots, H_K\}$, and $\overline{H}$ with random matrices
2.    Set $W^t = [W_1, \cdots, W_K]$ and $t = 1$
      Repeat
        For $k = 1, \cdots, K$
3.          Update $W_k^{t+1}$ with $W_k^{t+1} = W_k^t \circ \frac{V_k {H_k^t}^T + \lambda \overline{V} {\overline{H}_k^{t+1}}^T}{W_k^t H_k^t {H_k^t}^T + \lambda \overline{W}^t \overline{H}^t {\overline{H}_k^t}^T}$
4.          Update $H_k^{t+1}$ with $H_k^{t+1} = H_k^t \circ \frac{{W_k^{t+1}}^T V_k}{{W_k^{t+1}}^T W_k^{t+1} H_k^t}$
        End For
5.      Update $\overline{W}^{t+1} = [W_1^{t+1}, \cdots, W_k^{t+1}]$
6.      Split $\overline{H}^{t+1}$ into $\overline{H}^{t+1} = [{\overline{H}_1^{t+1}}^T, \cdots, {\overline{H}_k^{t+1}}^T]^T$
7.      Update $t \leftarrow t + 1$
      Until {The stopping condition (15) is satisfied.}
8.    Return $\{W_1, \cdots, W_K\}$, $\{H_1, \cdots, H_K\}$, and $\overline{H}$
```

TNMF provides a flexible framework for transductive NMF learning and various algorithms can be easily developed by replacing the Frobenious norm in (2) with other losses, e.g., Kullback Leibler divergence. **Algorithm 1** can be easily modified for optimizing TNMF variants and can be accelerated by utilizing the line search strategy introduced in [25,26]. In addition, the Frobenius norm based TNMF can be optimized by using the efficient NeNMF [33] method. We omit these studies due to the limit of space.

In summary, TNMF presents a friendly way of recognizing actions from still image due to the simplicity and flexibility of TNMF. We can easily construct the histograms of training actions and test image according to [4] and recognizing the action of the test image by the nearest neighbor (NN) classifier. By further incorporating constraints or regularizations on either features or encodings, interesting readers can easily extend this method for their own purposes in the future.

4 Experiments

Although the NMF-based method performs well on laboratory video frames [4], it is difficult to be applied to some tasks especially when some actions have insufficient examples, e.g., web images. This is because the pose clusters learned for some actions containing rare examples may be ill-posed.

4.1 Laboratory Datasets

For each collected images, we used an effective human detector [18] to detect people in different poses and aligned the detection rectangle by positioning the human head in its top-middle. Each of the detected human images is cropped and resized to a 78 × 42 color image. Based on the same image retrieval procedure for eight actions, we obtained a set of web images and extracted the HOG-descriptor for each cropped image. The HOG-descriptor for each image of each action is reshaped to a 1296-dimensional long vector and treated as a pose example [4].

Table 1. Statistics of the Google and Weizmann dataset, and ‘tr/ts’ means that the numbers of training poses and test poses are tr and ts, respectively

Action Name	‘run’	‘walk’	‘skip’	‘jump’	‘pjump’	‘wave’	‘jack’	‘bend’	‘side’
Google	201/202	285/286	67/68	118/119	109/109	52/53	43/44	30/30	-
Weizmann	30/165	129/238	30/184	30/140	103/167	283/326	90/206	97/84	96/124

Google Search Images. Figure 1(a) depicts some web images collected by using Google image search engine corresponding to human actions ‘run’, ‘walk’, ‘skip’, ‘jump’, ‘pjump’, ‘wave’, ‘jack’, and ‘bend’. For each action, e.g., ‘run’, we searched images on Google image search engine by using the keywords ‘run people’, ‘running people’, ‘run person’, and ‘running person’, and manually filtered all irrelevant images. Figure 1(b) shows the flow chart of generating the HOG descriptors of the Google web images. We constructed the Google dataset to include all the collected pose examples of web images.

Weizmann Video Frames. We conducted the same procedure on Weizmann video frames [1] which contains nine actions and formed another Weizmann pose dataset (or simply Weizmann dataset). Figure 2 gives examples of four actions including ‘run’, ‘walk’, ‘jump’, and ‘bend’ in the Weizmann dataset. It shows that video frames have more static backgrounds, and are therefore easier than the Google dataset.

Table 1 summarizes both datasets. It shows that actions ‘bend’ and ‘jack’ of the Google dataset contain a small number of training examples, and actions ‘run’, ‘skip’, and ‘jump’ of the Weizmann dataset contain a small number of training examples. Thus, the numbers of training examples for all actions are

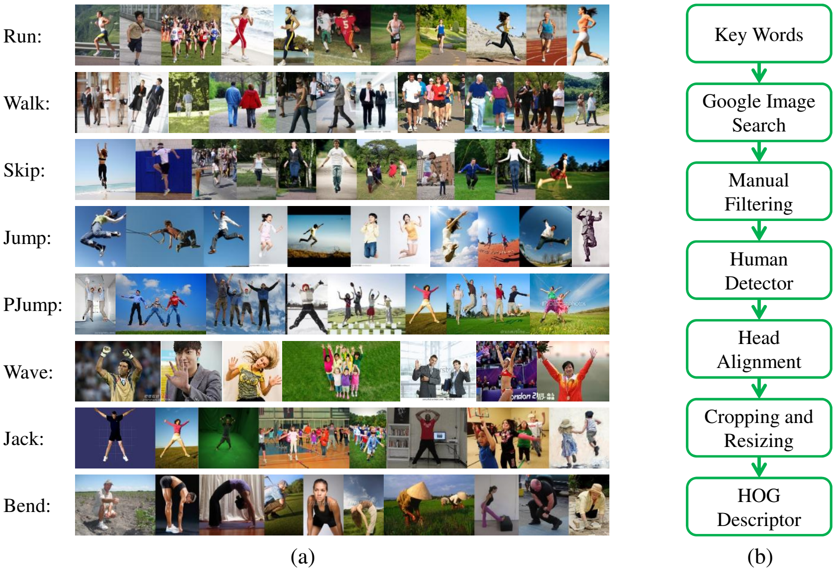

Fig. 1. Examples of web images returned by Google image search, where the action names from top to bottom are 'run', 'walk', 'skip', 'jump', 'pjump', 'wave', 'jack', and 'bend' (a), and (b) the flow chart of generating the HOG descriptor

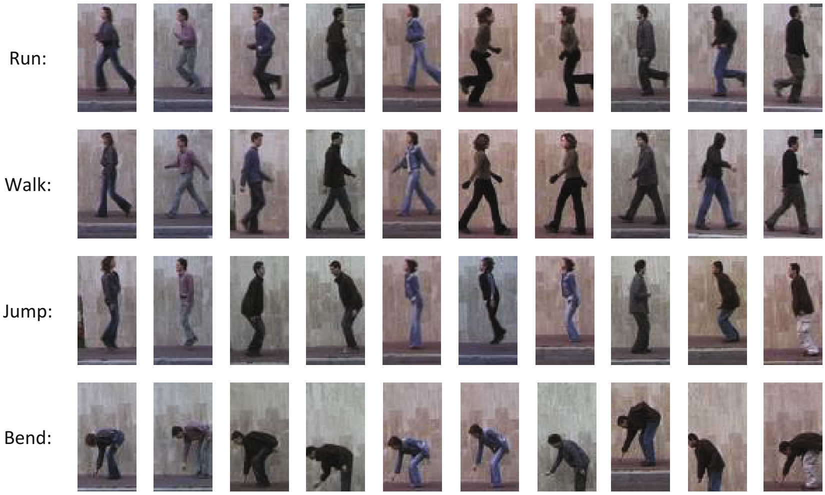

Fig. 2. Examples of video frames extracted from Weizmann dataset, where the action names from top to bottom are 'run', 'walk', 'jump', and 'bend'

imbalanced and performing NMF on the training examples of individual actions cannot obtain ‘effective’ primitive poses. In this experiment, we employed TNMF to overcome this deficiency by jointly learning dictionary from both training samples and test samples of each action. Although some actions have rare training examples, the dictionary obtained by simultaneously learning from both training and test samples are more discriminative than those obtained by separately learning from training samples [4]. To evaluate the effectiveness of TNMF, we compared the recognition accuracy of its learned dictionary with those learned by NMF.

According to [19], we first set the number of features for each action to 5 based on the number of common viewpoints for each action (2 for lateral views, 2 for views ±45° and 1 for frontal/back view), and cross-validated the trade-off parameter on a set $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Then we fixed the trade-off parameter to the best one, and cross-validated the number of features on a set $r \in \{5, 30, 50, 70, 90\}$. To evaluate the effectiveness of TNMF, Figure 3 gives the highest accuracies of NMF and TNMF obtained by cross-validation. Figure 3(a) and (b) show that TNMF outperforms NMF on Google dataset when varying r and λ in wide ranges of [50,90] and [0.1,0.7]. It shows that TNMF performs best when $\lambda = 0.7$ and $r = 50$. From Figure 3(c) and (d), we can see that MT-NMF outperforms NMF on Weizmann dataset when varying r and λ in wide ranges of [5,90] and [0.1,0.5], and it performs best when $\lambda = 0.1$ and $r = 70$.

Table 2. Accuracy (%) of NMF and TNMF on the Google and Weizmann dataset

Algorithms	NMF	TNMF
Google	74.66	**78.09**
Weizmann	88.30	**91.17**

Table 2 depicts the average accuracy of NMF and TNMF on both Google and Weizmann datasets. It shows that TNMF outperforms NMF on the Google dataset because it leverages the datasets across actions and learns better pose clusters for actions whose training examples are insufficient. The experimental results on the Weizmann dataset are consistent with this observation. It confirms the effectiveness of TNMF in action recognition from still images.

4.2 Willows Dataset

The Willow dataset[1] [22] contains totally 913 images for 7 activities including ‘interacting with computer’, ‘photographing’, ‘playing music’, ‘riding bike’, ‘riding horse’, ‘running’, and ‘walking’ (see Figure 4 for some examples of each action). Khan *et al.* [21] have demonstrated that fusing color and shape information can

[1] The Willow dataset is available at: `http://www.di.ens.fr/willow/research/stillactions/.`

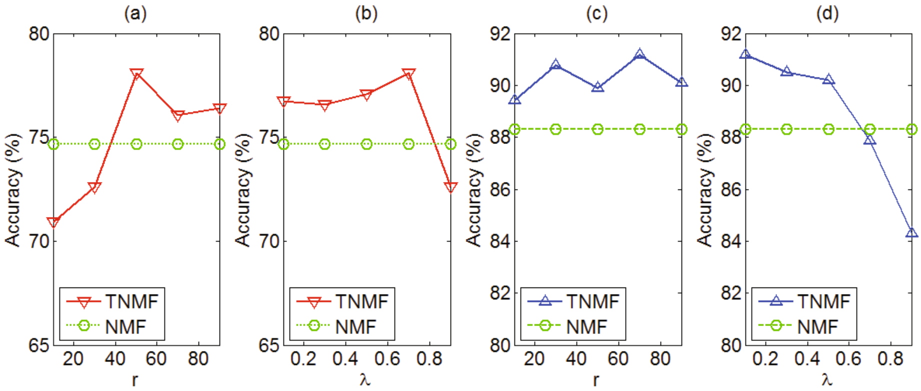

Fig. 3. Cross-validation of the number of features r and trade-off parameter λ of TNMF on the Google and Weizmann datasets, (a) accuracy versus r when $\lambda = 0.7$ and (b) accuracy versus λ when $r = 5$ on the Google dataset; (c) accuracy versus r when $\lambda = 0.1$ and (b) accuracy versus λ when $r = 70$ on the Weizmann dataset. The highest accuracies of NMF are included for comparison.

produce promising results of action recognition in still images. Along this direction, we extracted the shape cues by SIFT descriptors [32] and color cues by color names [34] separately, from each image, and fused them to construct the feature vector. The SIFT descriptor has 289 dimensionality and the color names has 11

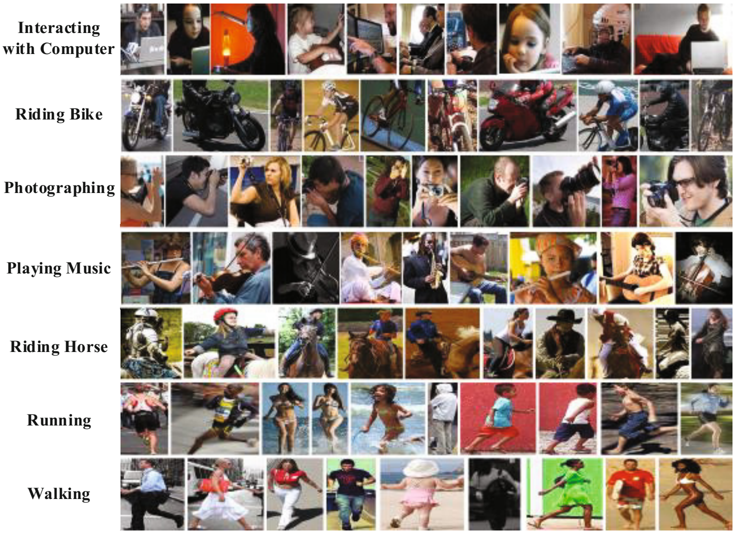

Fig. 4. Example images of different actions in the Willow dataset

dimensionality, and thus we constructed a 300-dimensional feature vector for each image.

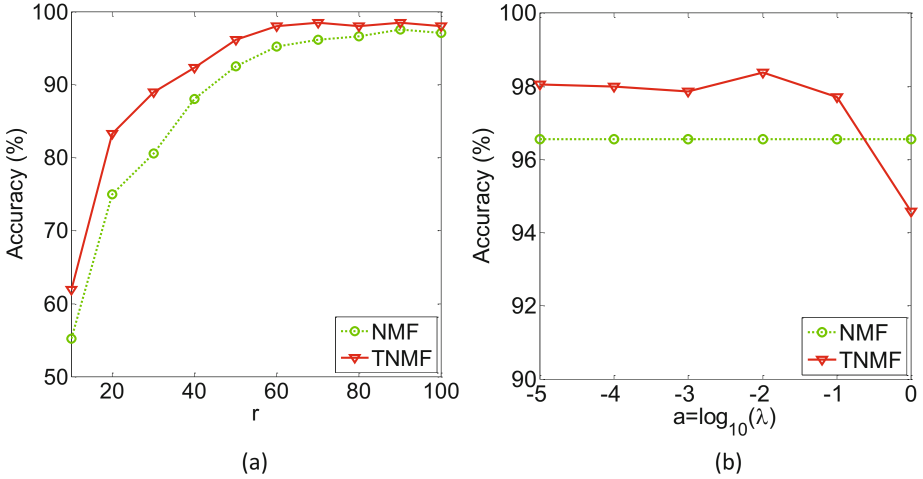

Fig. 5. Cross-validation of the reduced dimensionality r and trade-off parameter λ of TNMF on the Willow dataset: (a) accuracy when varying the reduced dimensionality r from 10 to 100 and fixing $\lambda = 0.1$, and the highest accuracy appears at $r = 70$; (b) accuracy when varying the trade-off parameter λ from 10^{-5} to 1 and fixing $r = 70$, and the highest accuracy appears at $\lambda = 10^{-2}$

In this experiment, we selected 100 images for each action, where 70 images are utilized for training and the remaining images are utilized for testing. To filter the influence of hyper-parameters of TNMF, i.e., the reduced dimensionality r and the trade-off parameter λ, to the final results, we varied the reduced dimensionality from 10 to 100 with a step size 10, and varied λ from 10^{-5} to 1. Such trial was repeated ten times for eliminate the influence of initialization of both TNMF and TNMF. Figure 5 shows that TNMF achieves the highest accuracy when $r = 70$ and $\lambda = 10^{-2}$, and that TNMF consistently outperforms NMF on the Willow dataset. This observation confirms the effectiveness of TNMF in still image based activity recognition.

4.3 Discussion

In summary, the experimental results on both laboratory dataset and real-world dataset demonstrate that the transductive learning trick in TNMF significantly improves the performance of action recognition still images. It should be honest that the TNMF based activity recognition method performs not very well when the number of actions are quite large, e.g., Stanford 40 [36] dataset. That is because the concatenation operator in TNMF (2) might lead to cancellation

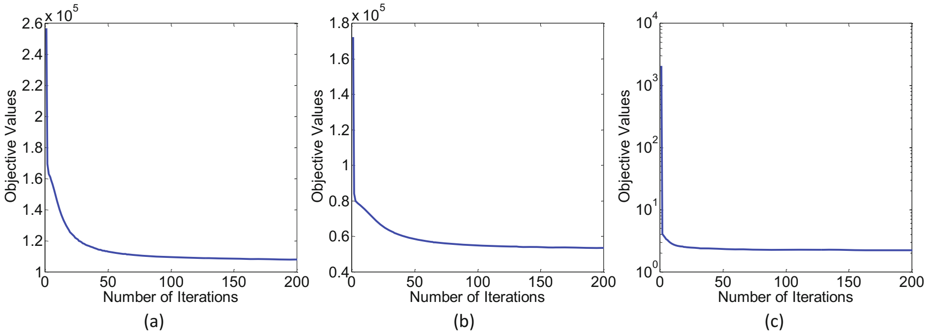

Fig. 6. The objective values versus number of iterations of TNMF on the Google (a), Weizmann (b), and Willow (c) datasets

among dictionaries of actions in this situation, and thus reduces the discriminative ability of the learned dictionary.

In this paper, we have theoretically proved the convergence of the MUR algorithm for TNMF. To verify this point, Figure 6 depicts the objective values versus number of iterations on the Google, Weizmann, and Willow datasets. They show that the MUR algorithm converges quite quickly, e.g., within 50 iteration rounds.

5 Conclusion

This paper proposes a novel method for activity recognition in still images based on transductive non-negative matrix factorization (TNMF). TNMF can transduce the visual features from training images to the learned encoding of test image. Therefore, TNMF boosts the performance of NMF based activity recognition especially on the datasets that contain insufficient training images for some actions. Experiments on both laboratory and real-world datasets demonstrate the effectiveness of TNMF.

Acknowledgments. This work is partially supported by Research Fund for the Doctoral Program of Higher Eduction of China, SRFDP (under grant No. 2034307110017) and Australian Research Council Projects (under grant No. FT-130101457 and DP-120103730).

References

1. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: International Conference on Computer Vision, vol. 2, pp. 1395–1402 (2005)
2. Laptev, I., Perez, P.: Retrieving actions in movies. In: Proceedings of International Conference on Computer Vision, pp. 1–8 (2007)

3. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. International Journal of Computer Vision **79**(3), 299–318 (2008)
4. Thurau, C., Hlavac, V.: Pose primitive based human action recognition in videos or still images. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
5. Aggarwal, J.K., Xia, L.: Human Activity Recognition from 3DData: A Review. Pattern Recognition Letters (2014)
6. Waltner, G., Mauthner, T., Bischof, H.: Indoor Activity Detection and Recognition for Sport Games Analysis. arXiv preprint arXiv:1404.6413 (2014)
7. Lee, D.D., Seung, H.S.: Learning the Parts of Objects with Non-negative Matrix Factorization. Nature **401**(6755), 788–791 (1999)
8. Xu, W., Liu, X., Gong, Y.: Document clustering based on nonnegative matrix factorization. In: ACM Special Interest Group on Information Retrieval, pp. 167–273 (2014)
9. Huang, X., Zheng, X., Yuan, W., Wang, F., Zhu, S.: Enhanced Clustering of Biomedical Documents Using Ensemble Nonnegative Matrix Factorization. Information Sciences **181**(11), 2293–2302 (2011)
10. Pauca, V., Piper, J., Plemmons, R.: Nonnegative Matrix Factorization for Spectral Data Analysis. Linear Algebra and its Applications **416**(1), 29–47 (2006)
11. Liu, L., Shao, L., Zhen, X., Li, X.: Learning Discriminative Key Poses for Action Recognition. IEEE Transactions on Cybernetics **43**(6), 1860–1870 (2013)
12. Zhang, Z., Tao, D.: Slow Feature Analysis for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3), 436–450 (2012)
13. Liu, L., Shao, L., Zheng, F., Li, X.: Realistic Action Recognition via Sparsely-constructed Gaussian Processes. Pattern Recognition **47**, 3819–3827 (2014)
14. Hotelling, H.: Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology **24**, 417–441 (1933)
15. Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics **7**, 179–188 (1936)
16. Guan, N., Lan, L., Tao, D., Luo, Z., Yang, X.: Transductive nonnegative matrix factorizationfor semi-supervised high-performance speech separation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2553–2557 (2014)
17. Lee, D.D., Seung, H.S.: Algorithms for Non-negative matrix factorization. In: Proceedings of Advances in Neural Information and Processing Systems, pp. 556–562 (2000)
18. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 7, pp. 1–8 (2008)
19. Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: Learning actions from the web. In: IEEE International Conference on Computer Vision, pp. 995–1002 (2009)
20. Zheng, Y., Zhang, Y.J., Li, X., Liu, B.D.: Action recognition in still images using a combination of human pose and context information. In: International Conference on Image Processing (2012)
21. Khan, F.S., Anwer, R.M., van deWeijer, J., Bagdanov, A.D., Lopez, A.M., Felsberg, M.: Coloring Action Recognition in Still Images. International Journal of Computer Vision **105**, 205–221 (2013)
22. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: astudy of bag-of-features and part-based representations. In: British Machine Vision Conference (2010)

23. Laptev, I.: On Space-time Interest Points. International Journal of Computer Vision **64**, 107–123 (2005)
24. Guo, G., Lai, A.: A Survey on Still Image Based Human Action Recognition. Pattern Recognition **47**, 3343–3361 (2014)
25. Guan, N., Tao, D., Luo, Z., Yuan, B.: Manifold Regularized Discriminative Non-negative Matrix Factorization with Fast Gradient Descent. IEEE Transactions on Image Processing **20**(7), 2030–2048 (2011)
26. Guan, N., Tao, D., Luo, Z., Yuan, B.: Non-negative Patch Alignment Framework. IEEE Transactions on Neural Networks **22**(8), 1218–1230 (2011)
27. Li, K., Fu, Y.: Prediction of Human Activity by Discovering Temporal Sequence Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(8), 1644–1657 (2014)
28. Kong, Y., Jia, Y., Fu, Y.: Interactive Phrases: Semantic Descriptions for Human Interaction Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(9), 1775–1788 (2014)
29. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6), 976–990 (2010)
30. Lambrecht, J., Kleinsorge, M., Rosenstrauch, M., Krger, J.: Spatial Programming for Industrial Robots Through Task Demonstration. International Journal of Advanced Robotic Systems 10(254) (2013)
31. Danafar, S., Gheissari, N.: Action recognition for surveillance applications using optic flow and SVM. In: Asian Conference on Computer Vision, pp. 457–466 (2007)
32. Lowe, D.G.: Distinctive Image Features from Scale-invariant Points. International Journal of Computer Vision **60**(2), 91–110 (2004)
33. Guan, N., Tao, D., Luo, Z., Yuan, B.: NeNMF: An Optimal Gradient Method for Non-negative Matrix Factorization. IEEE Transactions on Signal Processing **60**(6), 2882–2898 (2012)
34. van de Seijer, J., Schmid, C., Verbeek, J.J., Larlus, D.: Learning Color Names for Real-world Applications. IEEE Transactions on Image Processing **18**(7), 1512–1524 (2009)
35. Guo, G., Lai, A.: A survey on still image based human action recognition. Pattern Recognition **47**(10), 3343–3361 (2014)
36. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011