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

Naiyang Guan[1], Dacheng Tao[2], Long Lan[1], Zhigang Luo[1], and Xuejun Yang3
[1]Science and Technology on Parallel and Distributed Processing Laboratory, College of Computer Science, National University of Defense Technology, Changsha, Hunan, China (ny_guan@nudt.edu.cn, lan19901@126.com, zgluo@nudt.edu.cn)
[2]Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 235 Jones Street, Ultimo, NSW 2007, Australia (dacheng.tao@uts.edu.au)
[3]State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, Hunan, 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 HOG-descriptors of each human pose in the training images into fixed number of groups meanwhile learns to represent the HOG-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 original NMF. 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: Activity recognition on still images, 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, user interface design, robot trajectory planning, and surveillance thanks to the convenience of capturing videos through cameras [1][2][3]. 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.

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 [20] extracted from space-time volume, but the still image-based activity recognition usually cannot because only the spatial information is available on single images. The high-level cues can be characterized by various low-level features, e.g., color names [18], and different high-level cues can be combined to enhance the performance, e.g., combining pose and context information [17]. Interested readers can refer to [21] 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. Therefore, it is necessary to avoid explicitly reasoning about the body components [19]. In this paper, we constructed a high-level cues by clustering human poses with non-negative matrix factorization (NMF, [7]). Non-negative matrix factorization (NMF, [7]) 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, [11]) and Fisher's linear discriminative analysis (FLDA, [12]), 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][5][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 HOG-descriptors of the training poses of each action and the product of its features and encodings, and 2) it minimizes the distance between the HOG-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., $V \in R_+^{m\times n}$, NMF decomposes it into the product of two lower-rank matrices, i.e., $W \in R_+^{m\times r}$ and $H \in R_+^{r\times n}$, where $r \ll \min\{m,n\}$, by solving the following problem

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

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 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][9] and hyper-spectral imaging [10].

2.2 Transductive NMF

Recently, Guan *et al*. [13] 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

$$\min_{\forall 1\leq k\leq K, W_k\geq 0, H_k\geq 0, \bar{H}\geq 0} \left\{ \sum_{k=1}^{K} \|V_k - W_k H_k\|_F^2 + \lambda \|\bar{V} - \bar{W}\bar{H}\|_F^2 \right\}, \quad (2)$$

where $\bar{W} = [W_1, \ldots, W_K]$, 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 the parts-based representation of NMF, Thurau and Hlavac [4] proposed a static HOG-based NMF method for activity recognition on still images since the HOG-descriptor of 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 by

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

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

$$\bar{H} = \operatorname*{argmin}_{H\geq 0} \|\bar{V} - \bar{W}H\|_F^2, \quad (4)$$

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

At the classification stage, they calculated the histogram of each action based on $\{H_1, \ldots, H_K\}$, and the histogram of the test image based on $\bar{H}$, 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 on Still Images

In still image-based activity recognition, most actions have sufficient training images but some actions has rare images because the training images are widely collected from Internet and the activities are performed separately by different individuals. In this case, NMF cannot accurately

learn features on limited training images due to the overfitting problem.

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 from both training set and test set. 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 learn a dictionary on both training HOG-descriptors $\mathrm{V_k}$ from different actions and HOG-descriptors $\overline{\mathrm{V}}$ of the probe image. Since TNMF transduces the training poses to the learned dictionary by incorporating the second term in (4), it represents the probe poses more accurately and overcomes the deficiency of NMF. Experimental results confirm that TNMF greatly boosts the recognition performance.

Although the objective function of TNMF is jointly non-convex with respect to all variables $\{\mathrm{W}_1, \dots, \mathrm{W}_\mathrm{K}, \mathrm{H}_1, \dots, \mathrm{H}_\mathrm{K}, \overline{\mathrm{H}}\}$, it is convex with respect to each of them separately. According to [14], we utilized the majorization minimization (MM) method to derive a multiplicative update rule (MUR) for solving TNMF (2). MUR updates $\mathrm{W_k}$, $\mathrm{H_k}$, and $\overline{\mathrm{H}}$, respectively, by

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

$$\mathrm{H_k} \leftarrow \mathrm{H_k} \circ \frac{\mathrm{W_k^T V_k}}{\mathrm{W_k^T W_k H_k}}, \quad (6)$$

and

$$\overline{\mathrm{H}} \leftarrow \overline{\mathrm{H}} \circ \frac{\overline{\mathrm{W}}^\mathrm{T} \overline{\mathrm{V}}}{\overline{\mathrm{W}}^\mathrm{T} \overline{\mathrm{W}} \overline{\mathrm{H}}}, \quad (7)$$

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

$$\frac{|\mathrm{f}_t - \mathrm{f}_{t-1}|}{|\mathrm{f}_t - \mathrm{f}_0|} \leq \varepsilon, \quad (8)$$

where $\mathrm{f}_t = \sum_{\mathrm{k}=1}^{\mathrm{K}} \|\mathrm{V_k} - \mathrm{W}_k^t \mathrm{H}_k^t\|_\mathrm{F}^2 + \lambda \|\overline{\mathrm{V}} - \overline{\mathrm{W}}^t \overline{\mathrm{H}}^t\|_\mathrm{F}^2$ signifies the objective value at the t-th iteration round ($\mathrm{t} \geq 1$), 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: $\{V_1, \dots, V_1\}$, $\overline{\mathrm{V}}$, and r

Output: $\{\mathrm{W}_1, \dots, \mathrm{W}_\mathrm{K}\}$, $\{\mathrm{H}_1, \dots, \mathrm{H}_\mathrm{K}\}$, and $\overline{\mathrm{H}}$

1. Initialize $\{\mathrm{W}_1, \dots, \mathrm{W}_\mathrm{K}\}$, $\{\mathrm{H}_1, \dots, \mathrm{H}_\mathrm{K}\}$, and $\overline{\mathrm{H}}$ randomly
2. Set $\overline{\mathrm{W}} = [\mathrm{W}_1, \dots, \mathrm{W}_\mathrm{K}]$ and $\mathrm{t} = 1$

Repeat

For $\mathrm{k} = 1, \dots, \mathrm{K}$

3. Update W_k^{t+1} with $W_k^{t+1} = \mathrm{W}_k^t \circ \frac{\mathrm{V}_k \mathrm{H}_k^{t\,T} + \lambda \overline{\mathrm{V}} \overline{\mathrm{H}}_k^{t\,T}}{\mathrm{W}_k^t \mathrm{H}_k^t \mathrm{H}_k^{t\,T} + \lambda \overline{\mathrm{W}}^t \overline{\mathrm{H}}^t \overline{\mathrm{H}}_k^{t\,T}}$
4. Update H_k^{t+1} with $\mathrm{H}_k^{t+1} = \mathrm{H}_k^t \circ \frac{W_k^{t+1\,T} \mathrm{V_k}}{W_k^{t+1\,T} W_k^{t+1} \mathrm{H}_k^t}$

End For

5. Update $\overline{\mathrm{W}}^{t+1} = [\mathrm{W}_1^{t+1}, \dots, \mathrm{W}_K^{t+1}]$
6. Update $\overline{\mathrm{H}}^{t+1}$ with $\overline{\mathrm{H}}^{t+1} = \overline{\mathrm{H}}^t \circ \frac{\overline{\mathrm{W}}^{t+1\,T} \overline{\mathrm{V}}}{\overline{\mathrm{W}}^{t+1\,T} \overline{\mathrm{W}}^{t+1} \overline{\mathrm{H}}^t}$

7. Split $\bar{\mathrm{H}}^{t+1}$ into $\bar{\mathrm{H}}^{t+1} = \left[\bar{\mathrm{H}}_1^{t+1^T}, \ldots, \bar{\mathrm{H}}_K^{t+1^T}\right]^T$

8. Update $\mathrm{t} \leftarrow \mathrm{t}+1$

Until {The stopping condition (8) is satisfied}

9. **Return** $\{W_1^t, \ldots, W_K^t\}$, $\{H_1^t, \ldots, H_K^t\}$, and $\bar{\mathrm{H}}^t$

It is easy to verify that (6) and (7) decrease the objective function by using the auxiliary function technique in majorization minimization [14]. The auxiliary function is defined in **Definition 1** and has the property shown **Lemma 1**. By using the auxiliary function, **Proposition 1** proves that the multiplicative update rule (5) decreases the objective function. 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.

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

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

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

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

$$\mathrm{f}_t = \sum_{l \neq \mathrm{k}}^{\mathrm{K}} \|\mathrm{V}_l - \mathrm{W}_l^t \mathrm{H}_l^t\|_{\mathrm{F}}^2 + \|\mathrm{V}_k - \mathrm{W}_k \mathrm{H}_k^t\|_{\mathrm{F}}^2 + \lambda \|\bar{\mathrm{V}} - \bar{\mathrm{W}}^t \bar{\mathrm{H}}^t + W_k^t \bar{\mathrm{H}}_k^t - \mathrm{W}_k \bar{\mathrm{H}}_k^t\|_{\mathrm{F}}^2$$

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

$$\mathrm{f}(W_k) = \|\mathrm{V}_k - \mathrm{W}_k \mathrm{H}_k^t\|_{\mathrm{F}}^2 + \lambda \|\bar{\mathrm{V}} - \bar{\mathrm{W}}^t \bar{\mathrm{H}}^t + W_k^t \bar{\mathrm{H}}_k^t - \mathrm{W}_k \bar{\mathrm{H}}_k^t\|_{\mathrm{F}}^2. \quad (9)$$

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

$$\mathrm{g}(\mathrm{W}_k, \mathrm{W}_k^t) = \mathrm{f}(\mathrm{W}_k^t) + \langle \nabla \mathrm{f}(\mathrm{W}_k^t), \mathrm{W}_k - \mathrm{W}_k^t \rangle + \langle \frac{\mathrm{W}_k^t \mathrm{H}_k^t \mathrm{H}_k^{t^T} + \lambda \bar{\mathrm{W}}^t \bar{\mathrm{H}}^t \bar{\mathrm{H}}_k^{t^T}}{\mathrm{W}_k^t}, [\mathrm{W}_k - \mathrm{W}_k^t]^2 \rangle, \quad (10)$$

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

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

$$\begin{aligned} \mathrm{f}\left([\mathrm{W}_\mathrm{k}]_{ij}\right) = \mathrm{f}\left([\mathrm{W}_k^t]_{ij}\right) + [\nabla \mathrm{f}(\mathrm{W}_k^t)]_{ij}\left([\mathrm{W}_k]_{ij} - [\mathrm{W}_k^t]_{ij}\right) \\ + \left(\left[\mathrm{H}_k^t \mathrm{H}_k^{t^T}\right]_{jj} + \lambda\left[\bar{\mathrm{H}}_k^t \bar{\mathrm{H}}_k^{t^T}\right]_{jj}\right)\left([\mathrm{W}_k]_{ij} - [\mathrm{W}_k^t]_{ij}\right)^2. \quad (11) \end{aligned}$$

Since $\mathrm{H}_k^t \geq 0$ and $\mathrm{W}_k^t \geq 0$, we have

$$\left[\mathrm{H}_k^t \mathrm{H}_k^{t^T}\right]_{jj} \leq \frac{\sum_l [\mathrm{W}_k^t]_{il} \left[\mathrm{H}_k^t \mathrm{H}_k^{t^T}\right]_{lj}}{[\mathrm{W}_k^t]_{ij}} = \frac{\left[\mathrm{W}_k^t \mathrm{H}_k^t \mathrm{H}_k^{t^T}\right]_{ij}}{[\mathrm{W}_k^t]_{ij}}. \quad (12)$$

Since $\bar{\mathrm{H}}_k^t \geq 0$ and $\mathrm{W}_k^t \geq 0$, we have

$$\left[\bar{\mathrm{H}}_k^t \bar{\mathrm{H}}_k^{t^T}\right]_{jj} \leq \frac{\sum_l [\mathrm{W}_k^t]_{il} \left[\bar{\mathrm{H}}_k^t \bar{\mathrm{H}}_k^{t^T}\right]_{lj}}{[\mathrm{W}_k^t]_{ij}} = \frac{\left[\mathrm{W}_k^t \bar{\mathrm{H}}_k^t \bar{\mathrm{H}}_k^{t^T}\right]_{ij}}{[\mathrm{W}_k^t]_{ij}} \leq \frac{\left[\bar{\mathrm{W}}^t \bar{\mathrm{H}}^t \bar{\mathrm{H}}_k^{t^T}\right]_{ij}}{[\mathrm{W}_k^t]_{ij}}, \quad (13)$$

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

By substituting (12) and (13) into (11), we can easily verify that $\mathrm{f}(\mathrm{W_k}) \leq \mathrm{g}(\mathrm{W}_k, \mathrm{W}_k^t)$, and thus $\mathrm{g}(\mathrm{W}_k, \mathrm{W}_k^t)$ is an auxiliary function of $\mathrm{f}(\mathrm{W_k})$ according to **Definition 1**. By setting $\frac{\partial \mathrm{g}(\mathrm{W}_k, \mathrm{W}_k^t)}{\partial [\mathrm{W}_k]_{ij}} = 0$ and substituting $\nabla \mathrm{f}(\mathrm{W}_k^t) = (\mathrm{W}_k\mathrm{H}_k^t - \mathrm{V}_k)\mathrm{H}_k^{t\,T} + \lambda(\overline{\mathrm{W}}^t\overline{\mathrm{H}}^t - \overline{\mathrm{V}})\overline{\mathrm{H}}_k^{t\,T}$, we have

$$\left[\mathrm{W}_k^t\mathrm{H}_k^t\mathrm{H}_k^{t\,T}\right]_{ij} - \left[\mathrm{V}_k\mathrm{H}_k^{t\,T}\right]_{ij} + \lambda\left[\overline{\mathrm{W}}^t\overline{\mathrm{H}}^t\overline{\mathrm{H}}_k^{t\,T}\right]_{ij} - \lambda\left[\overline{\mathrm{V}}\overline{\mathrm{H}}_k^{t\,T}\right]_{ij} + \frac{\left[\mathrm{W}_k^t\mathrm{H}_k^t\mathrm{H}_k^{t\,T}\right]_{ij} + \lambda\left[\overline{\mathrm{W}}^t\overline{\mathrm{H}}^t\overline{\mathrm{H}}_k^{t\,T}\right]_{ij}}{[\mathrm{W}_k^t]_{ij}}\left([\mathrm{W}_k]_{ij} - [\mathrm{W}_k^t]_{ij}\right) = 0.$$

It is equivalent to

$$-\left[\mathrm{V}_k\mathrm{H}_k^{t\,T}\right]_{ij} - \lambda\left[\overline{\mathrm{V}}\overline{\mathrm{H}}_k^{t\,T}\right]_{ij} + \frac{\left[\mathrm{W}_k^t\mathrm{H}_k^t\mathrm{H}_k^{t\,T}\right]_{ij} + \lambda\left[\overline{\mathrm{W}}^t\overline{\mathrm{H}}^t\overline{\mathrm{H}}_k^{t\,T}\right]_{ij}}{[\mathrm{W}_k^t]_{ij}}[\mathrm{W}_k]_{ij} = 0. \quad (14)$$

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

$$[W_k^*]_{ij} = [\mathrm{W}_k^t]_{ij}\frac{\left[\mathrm{V}_k\mathrm{H}_k^{t\,T}\right]_{ij} + \lambda\left[\overline{\mathrm{V}}\overline{\mathrm{H}}_k^{t\,T}\right]_{ij}}{\left[\mathrm{W}_k^t\mathrm{H}_k^t\mathrm{H}_k^{t\,T}\right]_{ij} + \lambda\left[\overline{\mathrm{W}}^t\overline{\mathrm{H}}^t\overline{\mathrm{H}}_k^{t\,T}\right]_{ij}}. \quad (15)$$

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

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

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

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

$$\min_{\mathrm{W}_k \geq 0} \|\mathrm{X}_k - \mathrm{W}_k\mathrm{Y}_k\|_{\mathrm{F}}^2,$$

where $\mathrm{X}_k = \left[\mathrm{V}_k, \sqrt{\lambda}(\overline{\mathrm{V}} - \overline{\mathrm{W}}^t\overline{\mathrm{H}}^t + W_k^t\overline{\mathrm{H}}_k^t)\right]$ and $\mathrm{Y}_k = \left[\mathrm{H}_k^t, \sqrt{\lambda}\overline{\mathrm{H}}_k^t\right]$. It means that TNMF learns features both from training examples and test examples. In other words, TNMF achieves better generalization ability than NMF only on training examples.

TNMF provides a flexible framework for recognizing actions from still image due to its simplicity. By further incorporating constraints or regularizations on either features or encodings, interesting readers can easily extend it for their own purposes in the future works.

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 collected Internet. This is because the pose clusters learned for some actions containing rare examples may be ill-posed.

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.

Figure 1(a) depicts some web images 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. For each of the retrieved images, we used an effective human detector [15] 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. 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.

We conducted the same procedure on Weizmann video frames [1] which contains nine actions and formed another Weizmann pose dataset (or simply Weizmann dataset). Table I summarizes the two 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 imbalanced and performing NMF on the training examples of individual actions cannot obtain 'effective' primitive poses.

Table I. 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

In this experiment, we employed TNMF to overcome this deficiency by jointly learning features of all actions. We are motivated by the intuition that human poses of common actions such as 'run' and 'walk' are combined by some basic articulations. Although different actions have imbalanced training examples, primitive poses (bases) obtained by simultaneously learning from different actions are more effective than those obtained by separately learning from individual actions [4]. To evaluate the effectiveness of TNMF, we compared the recognition accuracy of its learned poses with those learned by NMF.

According to [16], 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 $\pm 45^{\circ}$ 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 2 gives the highest accuracies of NMF and TNMF obtained by cross-validation. Figure 2(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 2(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 $r = 70$.

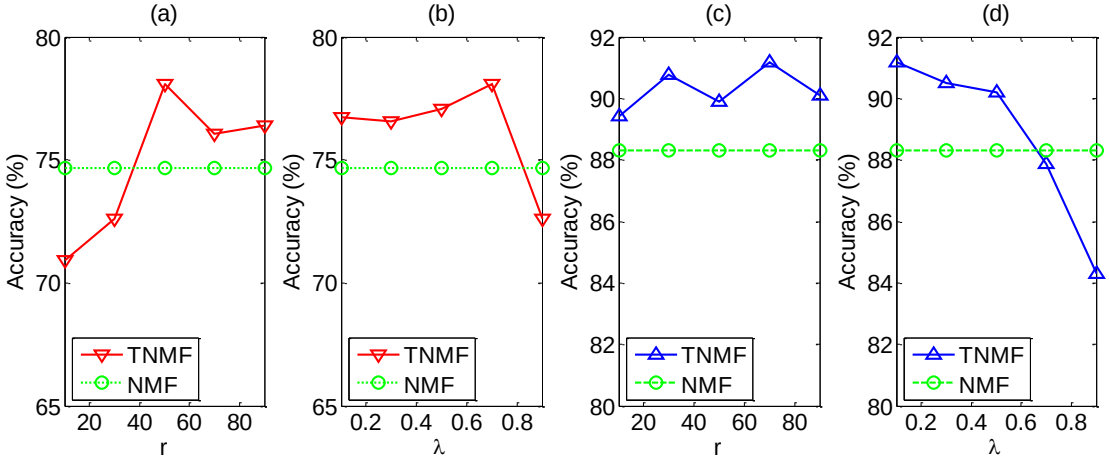

Fig. 2. Cross-validation of the number of features r and trade-off parameter λ of TNMF on the Google and Weizmann datasets, (a) accuracy versus r when $\lambda = 0.7$ and (b) accuracy versus λ when $r = 5$ on the Google dataset; (c) accuracy versus r when $\lambda = 0.1$ and (b) accuracy versus λ when $r = 70$ on the Weizmann dataset. The highest accuracies of NMF are included for comparison.

Table II 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 training examples and test examples, 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 on still images.

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

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

In summary, the experimental results of both laboratory dataset and real-world dataset demonstrate that the transductive learning trick in TNMF significantly improves the performance

of action recognition on still images.

5. Conclusion

This paper proposes a novel method for activity recognition on still images called transductive non-negative matrix factorization (TNMF). TNMF can transduce the visual features from training HOG-descriptors to the learned encoding of test image. Therefore, TNMF boosts the performance of activity recognition especially on the datasets that contain imbalanced number of images among different actions. Experiments on both laboratory and real-world datasets demonstrate that TNMF consistently improves the performance of NMF.

Acknowledgement

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] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. *In International Conference on Computer Vision*, vol. 2, pp. 1395-1402, 2005.

[2] Laptev and P. Perez. Retrieving Actions in Movies. *In Proceedings of International Conference on Computer Vision*, pp. 1-8, 2007.

[3] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299-318, 2008.

[4] C. Thurau and V. Hlavac. Pose Primitive based Human Action Recognition in Videos or Still Images. *In IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[5] J.K. Aggarwal and L. Xia. Human Activity Recognition from 3D Data: A Review. *Pattern Recognition Letters*, 2014.

[6] G. Waltner, T. Mauthner, and H. Bischof. Indoor Activity Detection and Recognition for Sport Games Analysis. *arXiv preprint arXiv:1404.6413*, 2014.

[7] D.D. Lee and H.S. Seung. Learning the Parts of Objects with Non-negative Matrix Factorization. *Nature*, vol. 401, no. 6755, pp. 788-791, 1999.

[8] W. Xu, X. Liu, and Y. Gong. Document Clustering Based on Nonnegative Matrix Factorization. *In ACM Special Interest Group on Information Retrieval*, pp. 167-273, 2003.

[9] X. Huang, X. Zheng, W. Yuan, F. Wang, and S. Zhu. Enhanced Clustering of Biomedical Documents Using Ensemble Nonnegative Matrix Factorization, *Information Sciences*, vol. 181, no. 11, pp. 2293–2302, Jun. 2011.

[10] V. Pauca, J. Piper, and R. Plemmons. Nonnegative Matrix Factorization for Spectral Data Analysis, *Linear Algebra and its Applications*, vol. 416, no. 1, pp. 29-47, Jul. 2006.

[11] H. Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, vol. 24, pp. 417-441, 1933.

[12] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[13] N. Guan, L. Lan, D. Tao, Z. Luo, and X. Yang. Transductive Nonnegative Matrix Factorization for Semi-supervised High-performance Speech Separation. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2553-2557, 2014.

[14] D.D. Lee and H.S. Seung. Algorithms for Non-negative Matrix Factorization. *In Proceedings*

of Advances in Neural Information and Processing Systems, pp. 556-562, 2000.

[15] P. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.

[16] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff. Learning Actions From the Web. *In IEEE International Conference on Computer Vision*, pp. 995-1002. 2009.

[17] Y. Zheng, Y.J. Zhang, X. Li, and B.D. Liu. Action Recognition in Still Images Using a Combination of Human Pose and Context Information. *In International Conference on Image Processing*, 2012.

[18] F.S. Khan, R.M. Anwer, J. van deWeijer, A.D. Bagdanov, A.M. Lopez, and M. Felsberg. Coloring Action Recognition in Still Images. *International Journal of Computer Vision*, vol. 105, pp. 205-221, 2013.

[19] V. Delaitre, I. Laptev, and J. Sivic. Recognizing Human Actions in Still Images: A Study of Bag-of-features and Part-based Representations. *British Machine Vision Conference*, 2010.

[20] I. Laptev. On Space-time Interest Points. *International Journal of Computer Vision*. vol. 64, pp. 107-123, 2005.

[21] G. Guo and A. Lai. A Survey on Still Image Based Human Action Recognition. *Pattern Recognition*, vol. 47, pp. 3343-3361, 2014.