RECOGNIZING HUMAN ACTIONS FROM LOW-RESOLUTION VIDEOS BY REGION-BASED MIXTURE MODELS

Ying Zhao[1,2,3], *Huijun Di*[1], *Jian Zhang*[2], *Yao Lu*[1*], *Feng Lv*[1]

[1]Beijing Laboratory of Intelligent Information Technology,Beijing Institute of Technology, Beijing, China
[2]Advanced Analytics Institute, University of Technology, Sydney, Australia
[3]Teachers College, Beijing Union University, Beijing, China
sftzhaoying@buu.edu.cn, {ajon,vis_yl,lvfeng}@bit.edu.cn, jian.zhang@uts.edu.au

ABSTRACT

Recognizing human action from low-resolution (LR) videos is essential for many applications including large-scale video surveillance, sports video analysis and intelligent aerial vehicles. Currently, state-of-the-art performance in action recognition is achieved by the use of dense trajectories which are extracted by optical flow algorithms. However, the optical flow algorithms are far from perfect in LR videos. In addition, the spatial and temporal layout of features is a powerful cue for action discrimination. While, most existing methods encode the layout by previously segmenting body parts which is not feasible in LR videos. Addressing the problems, we adopt the Layered Elastic Motion Tracking (LEMT) method to extract a set of long-term motion trajectories and a long-term common shape from each video sequence, where the extracted trajectories are much denser than those of sparse interest points(SIPs); then we present a hybrid feature representation to integrate both of the shape and motion features; and finally we propose a Region-based Mixture Model (RMM) to be utilized for action classification. The RMM models the spatial layout of features without any needs of body parts segmentation. Experiments are conducted on two publicly available LR human action datasets. Among which, the UT-Tower dataset is very challenging because the average height of human figures is only about 20 pixels. The proposed approach attains near-perfect accuracy on both of the datasets.

Index Terms— Low-resolution(LR), Action Recognition, Elastic Motion Tracking, Mixture Model

1. INTRODUCTION

Human action recognition from low-resolution (LR) videos typically happens when the videos are taken from a far field of view, where the resolution of human figure is very low (generally ranging from 20 to 50 pixels in height), as illustrated in Fig. 1. Recognizing human actions accurately from such LR videos is essential for many applications including large-scale video surveillance, sports video analysis and intelligent aerial vehicles.

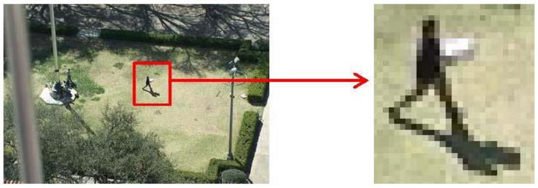

Fig. 1. A Sample Frame from *UT-Tower* Dataset. The human figure in the frame is only about 20 pixels in height.

LR human action recognition suffers more challenges than those presented in medium or high-resolution videos. First, the appearance of human figure is usually blurry and the configuration of body parts tends to be barely distinguishable, which means that exact description of human appearance and accurate segmentation of body parts are not feasible in LR scenarios. Second, the size of human figure is too limited in LR videos, where the average width of human limbs is only about 2 or 3 pixels, and the state-of-the-art optical flow algorithms are far from perfect under such too limited size[1, 2], so a more accurate and robust tracking method should be adopted to extract reliable motion features for LR human action recognition.

Addressing the challenges, we propose to adopt the Layered Elastic Motion Tracking (LEMT) method [3, 4] to track the human action. As a result, a long-term common shape and a set of long-term motion trajectories are extracted from each video sequence, as shown in Fig. 2. We further present a hybrid feature representation to integrate the extracted shape and motion features, and propose a Region-based Mixture Model (RMM) for action recogintion.

The main contributions of the paper are two-fold.

(1) We present a hybrid feature representation to effec-

[*]Corresponding author is Yao Lu. Email address: vis_yl@bit.edu.cn. This work was supported in part by the 2014 General/Senior Foreign Visiting Scholars Training Project for Excellent Young Teachers in Beijing Colleges Governed by the Beijing Municipal Commission of Education (Project number: 067145301400), by the National Natural Science Foundation of China (No.61273273), and by the Research Fund for the Doctoral Program of Higher Education of China (No. 20121101110034).

tively integrate the shape and motion features. Since optical flow algorithms are far from perfect in LR videos, we present an alternative approach to extract reliable shape and motion features from each video sequence by adopting the LEMT method; and further we present a hybrid feature representation method to effectively integrate the extracted shape and motion features.

(2) We learn a RMM for each action category to be utilized for action recognition. The RMMs are more informative than pure bag-of-features (BOF) methods by considering the spatial layout of features and are able to overcome the difficulties of body parts segmentation in LR videos.

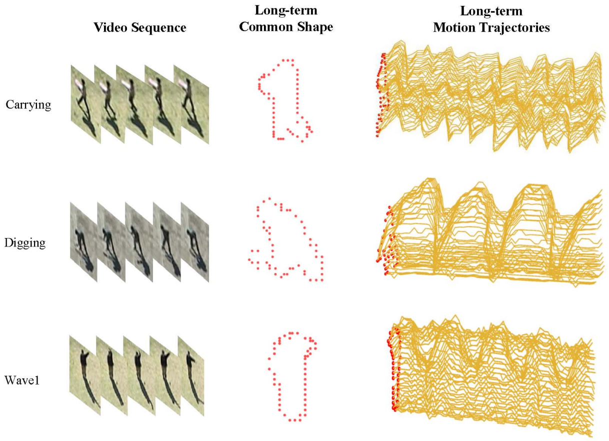

Fig. 2. Long-term Common Shapes and Long-term Motion Trajectories of Some Actions in *UT-Tower* Dataset. Each point in the long-term common shape has a corresponding long-term motion trajectory.

2. RELATED WORKS

The paper by Efros et al. [5] is one of the earliest works on human action recognition in LR videos. They presented a motion descriptor to represent an action and adopted the k-nearest-neighbor classifier to recognize actions. However, the only use of motion feature is insufficient to discriminate some “static” actions, such as *pointing* and *standing*.

Chen et al. [2] combined both human poses and motion information to characterize human actions and they used Support Vector Machine (SVM) to classify actions.They also created and published a UT-Tower dataset which has been a standard dataset for researchers to evaluate the performance of LR human action recognition methods nowadays.

Later on, Ryoo et al. [6] held the “Aerial View Activity Classification Challenge” around the world to encourage researchers to explore techniques for accurately recognizing human actions in LR videos. There were 4 university teams who participated in the challenge and the UT-Tower dataset was used to evaluate each participant’s method. The winner is the team from the Boston University [7].

Recently, state-of-the-art performance in action classification is achieved by extracting dense trajectories. Wang et al. [1] achieved a significantly improved performance in action recognition by the use of dense trajectories which were extracted by computing dense optical flow field. However, the state-of-the-art optical flow algorithms are far from perfect in LR videos [1, 2].

The spatial and temporal layout of features is a powerful cue for action discrimination. Currently, Ciptadi et al. [8] proposed a movement pattern histogram (MPH) to encode the temporal layout of features by decomposing a human action into several movement primitives (corresponding roughly to body parts). While,it is not a reasonable hope of decomposing body parts in LR videos.

In this paper, we extract both of the shape and motion features to represent human action informatively and discriminatively. Particularly, our motion trajectories are extracted by the LEMT method [3, 4] instead of optical flow algorithms and are much denser than those of SIPs. Besides, we learn a RMM for each action category, which models the spatial layout of features without any needs of body parts segmentation.

3. FEATURE EXTRACTION AND REPRESENTATION

3.1. Feature Extraction

Given a video sequence, we firstly localize the region of interest (ROI) in each frame and extract the raw edges of human body in each ROI. In LR videos, one of the relatively reliable visual cues is the human shape which could be represented by edges of human body. Edges of human body could be extracted at very low cost when the foreground blob is available. By following other published papers, such as [7, 9, 10, 11, 12], we assume that the foreground masks of each video frame are available to us so that we could focus on the recognition problem instead of foreground segmentation issue. By the use of pixel corresponding relationships between the original frames and the foreground masks, we get the foreground blobs; and then we extract the raw edges of human body in each frame by applying the edge detection method[13] on the foreground blobs.

We further apply the LEMT method [3, 4] on the raw edges of human body to extract a long-term common shape and a set of long-term motion trajectories from each video sequence (as shown in Fig. 2), where the former reflects the long-term stable geometric structure of the whole human body when performing the action and the latter ensures a good coverage of human motion because they are much denser than those of SIPs.

3.2. Feature representation

We integrate features in the long-term common shape and in the long-term motion trajectories into a hybrid feature set. Firstly, we normalize all the long-term common shapes and long-term motion trajectories. Secondly, we represent each long-term common shape as a point set $\{(x_i^0, y_i^0); i = 1, 2, ..., N\}$, where N is the number of points in the long-term common shape, and (x_i^0, y_i^0) denotes the position of the ith point in the long-term common shape. Thirdly, we calculate a motion vector from each long-term motion trajectory. The long-term motion trajectory extracted by the LEMT method [3, 4] is represented by a sequence of coordinate values $\{(x_i^t, y_i^t); i = 1, 2, ..., N; t = 1, 2, ..., T\}$, where T is the number of frames in the video sequence, and (x_i^t, y_i^t) denotes the position of the ith point at frame t. We calculate a motion vector $(\mu_i^x, \mu_i^y, e_i^x, e_i^y, v_i^x, v_i^y)^T$ from each of the trajectories, where μ_i^x and μ_i^y are separately the mean of x-coordinate values and the mean of y-coordinate values of the ith trajectory, e_i^x and e_i^y separately denote the motion energy in x-direction and in y-direction of the ith trajectory, v_i^x and v_i^y separately denote the average moving velocity in x-direction and in y-direction of the ith trajectory. The e_i^x, e_i^y, v_i^x and v_i^y are respectively calculated by formula (1), (2), (3) and (4). Finally, we combine these two kinds of features into a hybrid feature set $W = \{w_i; i = 1, 2, ..., N\}$, where $w_i = \{g_i, m_i\}$ denotes the position vector $g_i = (x_i^0, y_i^0)^T$ of the ith point in the long-term common shape and the motion vector $m_i = (\mu_i^x, \mu_i^y, e_i^x, e_i^y, v_i^x, v_i^y)^T$ of the ith long-term motion trajectory.

$$e_i^x = \sqrt{\frac{\sum_{t=1}^{T}(x_i^t - \mu_i^x)^2}{T-1}} \tag{1}$$

$$e_i^y = \sqrt{\frac{\sum_{t=1}^{T}(y_i^t - \mu_i^y)^2}{T-1}} \tag{2}$$

$$v_i^x = \frac{\sum_{t=1}^{T-1}|x_i^{t+1} - x_i^t|}{T-1} \tag{3}$$

$$v_i^y = \frac{\sum_{t=1}^{T-1}|y_i^{t+1} - y_i^t|}{T-1} \tag{4}$$

4. ACTION MODELING AND RECOGNITION

Human actions are results of body parts movements. It is reasonable for one to model an action as a combination of body parts movements, where each body part has its stable geometric structure and motion pattern. However, it is not feasible to accurately separate body parts in LR videos. To overcome this difficulty, we propose the RMM.

4.1. The Region-based Mixture Model

As illustrated in Fig. 3, the RMM is composed of K components which are mixed together by K mixture coefficients. Each component corresponds to a stable and identifiable region in the long-term common shapes of the action category with its own shape and motion distributions. Each hybrid feature is softly assigned to each region according to the probability that it is generated from the shape and motion distributions of the region.

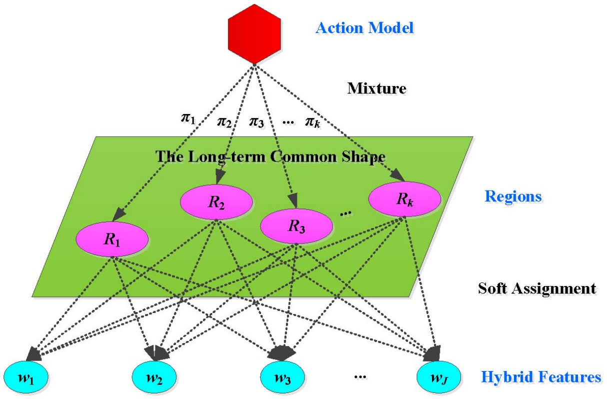

Fig. 3. Illustration of the RMM for an Action Category. The spatial layout of features is modeled by the stable and identifiable regions in the long-term common shapes of the action category. Each region has its shape and motion distributions.

The RMM is parameterized by $\Theta = \{\pi_k, \theta_k^G, \theta_k^M; k = 1, 2, ..., K\}$, where K is the number of regions, π_k is the mixture coefficient of the kth region and $\sum_{k=1}^{K}\pi_k = 1$, θ_k^G are parameters that govern the kth region shape distribution, θ_k^M are parameters that govern the kth region motion distribution. Each hybrid feature w_i is viewed as a random sample generated from the RMM, and the density for each w_i is

$$p(w_i|\Theta) = \sum_{k=1}^{K}\pi_k p(g_i|\theta_k^G)p(m_i|\theta_k^M) \tag{5}$$

In practice, we model the shape distribution of a region as a Gaussian distribution which is governed by parameters $\theta_k^G = \{\mu_k^G, \Sigma_k^G\}$, and model the motion distribution of a region also as a Gaussian distribution which is govern by parameters $\theta_k^M = \{\mu_k^M, \Sigma_k^M\}$.

4.2. Learning Model Parameters

Given L training video sequences of an action category, we extract L hybrid feature sets $Z = \{W_l; l = 1, 2, ..., L\}$, where W_l is the hybrid feature set that exacted from the lth training video sequence. The problem of learning model parameters is actually a maximum likelihood estimation (MLE) problem which is to find the best Θ maximizing the log-likelihood $log(p(Z|\Theta))$.We adopt expectation maximization (EM) algorithm to solve the problem.

Initialization: We build a compact vector $v_j = (x_j^0, y_j^0, \mu_j^x, \mu_j^y, e_j^x, e_j^y, v_j^x, v_j^y)^T$ by concatenating the position vector $g_j = (x_j^0, y_j^0)^T$and the motion vector $m_j = (\mu_j^x, \mu_j^y, e_j^x, e_j^y, v_j^x, v_j^y)^T$ in each hybrid feature w_j, where $j = 1, 2, ..., J$ and J is the total number of hybrid features in Z. Then, we adopt k-means algorithm to cluster all the compact vectors $\{v_j; j = 1, 2, ..., J\}$ into K clusters. Inside each cluster, we compute the initial values $\mu_k^{G(0)}$ and $\Sigma_k^{G(0)}$ with the use of $\{(x_s^0, y_s^0)^T \in \text{cluster } k; s = 1, 2, ...J_k\}$, the initial values $\mu_k^{M(0)}$ and $\Sigma_k^{M(0)}$ with the use of $\{(\mu_s^x, \mu_s^y, e_s^x, e_s^y, v_s^x, v_s^y)^T \in \text{cluster } k; s = 1, 2, ...J_k\}$, and the initial occupancy probability of each resion $\pi_k^{(0)} = J_k/J$, where J_k is the number of features belonging to the kth cluster. Finally, we get the initial estimate of the model parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \mu_k^{G(0)}, \Sigma_k^{G(0)}, \mu_k^{M(0)}, \Sigma_k^{M(0)}; k = 1, 2, ..., K\}$.

E-Step: we evaluate the expected value as follow

$$
\begin{aligned}
Q(\Theta, \Theta^{(0)}) = & \sum_{k=1}^{K}\sum_{j=1}^{J} log(\pi_k) p(k|w_j, \Theta^{(0)}) \\
& + \sum_{k=1}^{K}\sum_{j=1}^{J} log(\mathcal{N}(g_i|\mu_k^G, \Sigma_k^G)) p(k|w_j, \Theta^{(0)}) \\
& + \sum_{k=1}^{K}\sum_{j=1}^{J} log(\mathcal{N}(m_j|\mu_k^M, \Sigma_k^M)) p(k|w_j, \Theta^{(0)})
\end{aligned} \tag{6}
$$

where

$$p(k|w_j, \Theta^{(0)}) = \frac{\pi_k^{(0)} p(w_j|\theta_k^{(0)})}{\sum_{r=1}^{K} \pi_r^{(0)} p(w_j|\theta_r^{(0)})} \tag{7}$$

and

$$p(w_j|\theta_k^{(0)}) = \mathcal{N}(g_j|\mu_k^{G(0)}, \Sigma_k^{G(0)}) \mathcal{N}(m_j|\mu_k^{M(0)}, \Sigma_k^{M(0)}) \tag{8}$$

M-Step: By maximizing the expectation which isd by (6), We find the updated expressions for each parameter:

$$\pi_k^* = \frac{1}{J}\sum_{j=1}^{J} p(k|w_j, \Theta^{(0)}) \tag{9}$$

$$\mu_k^{G^*} = \frac{\sum_{j=1}^{J} g_j p(k|w_j, \Theta^{(0)})}{\sum_{j=1}^{J} p(k|w_j, \Theta^{(0)})} \tag{10}$$

$$\Sigma_k^{G^*} = \frac{\sum_{j=1}^{J} p(k|w_j, \Theta^{(0)})(g_j - \mu_k^G)(g_j - \mu_k^G)^T}{\sum_{j=1}^{J} p(k|w_j, \Theta^{(0)})} \tag{11}$$

$$\mu_k^{M^*} = \frac{\sum_{j=1}^{J} m_j p(k|w_j, \Theta^{(0)})}{\sum_{j=1}^{J} p(k|w_j, \Theta^{(0)})} \tag{12}$$

$$\Sigma_k^{M^*} = \frac{\sum_{j=1}^{J} p(k|w_j, \Theta^{(0)})(m_j - \mu_k^M)(m_j - \mu_k^M)^T}{\sum_{j=1}^{J} p(k|w_j, \Theta^{(0)})} \tag{13}$$

where $\Sigma_k^{M^*}$ and $\Sigma_k^{G^*}$ are diagonal covariance matrices.

4.3. Action Recognition

With the learned RMMs which are parameterized by Θ^c, where $c = 1, 2, ..., C$ and C is the number of RMMs (i.e. the number of action categories), the task of recognizing a new input action is to find a RMM which best matches the hybrid feature set W^{new} extracted from the new input video sequence in the same way as that of training video sequences. We find the best matching RMM by calculating and comparing the posterior distribution $p(\Theta^c|W^{new})$ of each RMM. The posterior probability of each RMM is caculated by

$$p(\Theta^c|W^{new}) \propto \frac{N_c}{N_{Total}} p(W^{new}|\Theta^c) \tag{14}$$

where N_c is the number of training samples for the cth action category, N_{Total} is the total number of training samples for all action categories, $p(W^{new}|\Theta^c) = \prod_{i=1}^{N^{new}} p(w_i^{new}|\Theta^c)$and N^{new} is the number of hybrid features in W^{new}.

5. EXPERIMENTS AND RESULTS

We evaluate the performance of the proposed approach on two publicly available LR datasets: the Weizmann dataset [9] and the UT-Tower dataset [14], both of which provide the foreground masks. Even though there exist other public human action datasets, such as the Hollywood dataset and the UCF101 dataset, we omit them in our experiments because they have human figures in medium or high-resolution. Although there are other LR datasets, for example the Soccer dataset and the VIRAT dataset, we do not use them in our experiments. Because, as declared in section 3.1, we focus on the recognition problem instead of foreground segmentation issue in this paper. By following other published papers [7, 9, 10, 11, 12], we assume that the foreground masks are provided by datasets so that we could directly use them as inputs of our experiments. While neither Soccer nor VIRAT dataset provides these masks.

For evaluation, we adopt the standard protocol of these two dataset: leave-one-out cross validation (LOOCV) scheme. Experimental results and comparisons show that our approach is comparable to or better than state-of-the-art results and, more importantly, our approach works directly with the raw edges of the human body without any needs of body parts segmentation, which makes it more general for LR action recognition tasks.

5.1. Weizmann dataset

There are 10 types of actions in Weizmann dataset, which are *bend*, *run*, *skip*, *jack*, *jump*, *pjump*, *side*, *wave1*, *wave2* and *walk*. In implementation, we use a fixed rectangular as a ROI for each frame, since the view is static in the dataset. We learn a RMM for each action category. The learned RMMs imposed over testing frames are shown in Fig. 4. The number of regions for each RMM is adjusted manually to get the best recognition accuracy. That is, we learn 8-regions RMMs for *walking* and *running*; 7-regions RMMs for *skipping*, *jacking* and *siding*; 6-regions RMM for *wave2*; 5-regions RMMs for *jumping*, *wave1* and *bending*; 3-regions RMM for *pjumping*. Table 1 shows that our approach is comparable to or better than state-of-the-art results on the Weizmann dataset.

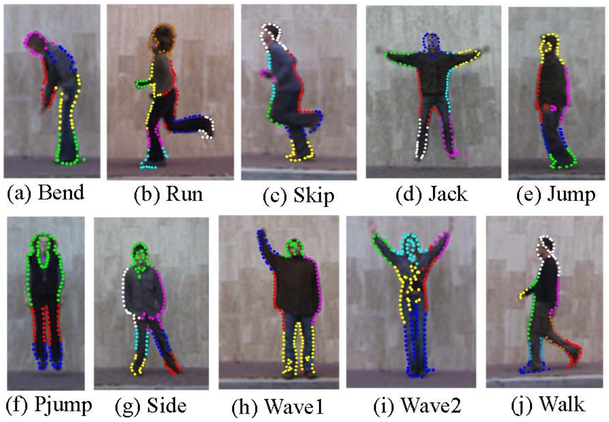
(a) Bend (b) Run (c) Skip (d) Jack (e) Jump
(f) Pjump (g) Side (h) Wave1 (i) Wave2 (j) Walk

Fig. 4. Visualization of the learned RMMs imposed over testing frames on the *Weizmann* dataset. Regions are presented in different color. Some interesting observations can be made. For example, in Fig. 4(a), the pink region seems to correspond roughly to the "head"; the red region roughly to the "arm"; the yellow region roughly to the "upper legs"; the green region roughly to the "lower legs"; and the navy blue region roughly to the "back".

Table 1. Comparisons of Average Accuracy on *Weizmann*

Method	Average Accuracy (%)
Hejin Yuan, 2015 [10]	93.55
Yang Y *et al.*, 2013 [15]	99
Chaaraoui A A *et al.*, 2013 [11]	92.77
Reddy K K et al., 2012 [16]	90.32
The Proposed Approach	**98.92**

5.2. UT-Tower dataset

To show the effectiveness of our method on more challenging LR scenarios, we evaluate it on the UT-Tower dataset which was created to simulate aerial view video surveillance by taking videos from the top of the 307-foot tall UT Austin Tower building. The average height of human figures in this dataset is only about 20 pixels. In addition to the LR setup, the UT-Tower dataset also poses other challenges. For example, the direct sunlight causes salient human cast shadows.

There are 9 categories of human actions in the dataset, which are *pointing*, *standing*, *digging*, *walking*, *carrying*, *running*, *wave1*, *wave2* and *jumping*. In implementation, we use the provided bounding boxes as a ROI for each frame, and remove the shadows of the foreground blobs inside each ROI by the use of shadow removal method [17].

We learn 9 RMMs for the 9 action categories. The Visualization of the learned RMMs imposed over testing frames is shown in Fig. 5. The number of regions for each RMM is also adjusted manually to get the best recognition accuracy. That is, we learn 2-regions RMMs for *pointing*, *standing* and *jumping*; 3-regions RMMs for *walking* and *running*; 4-regions RMMs for *digging* and *wave1*; 6-regions RMMs for *carrying* and *wave2*. Comparing to the state-of-the-art methods, as shown in Table 2, we obtained the best result with a simpler approach. Our approach is able to work without any needs of previously separating body parts, which makes it more general for LR application.

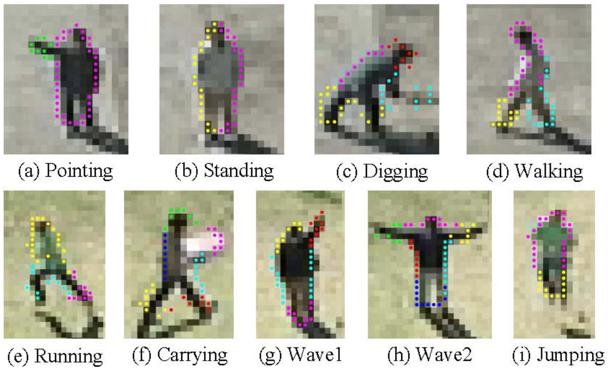
(a) Pointing (b) Standing (c) Digging (d) Walking
(e) Running (f) Carrying (g) Wave1 (h) Wave2 (i) Jumping

Fig. 5. Visualization of the learned RMMs imposed over testing frames on the *UT-Tower* dataset. Regions are presented in different color. Similarly, we can also make some interesting observations. For example, in Fig. 5(a), the green region seems to correspond roughly to the "pointing arm" and the pink region roughly to the "torso" of the human body.

Table 2. Comparisons of Average accuracy on *UT-Tower*

Method	Average Accuracy (%)
Zhao K *et al.*, 2015 [18]	92.45
Guo K *et al.*, 2013 [12]	97.22
Cao X et al., 2012 [19]	98.15
Mukherjee S *et al.*, 2011 [20]	97.22
The Proposed Approach	**99.07**

6. CONCLUSIONS

Human action recognition in LR videos is an important but challenging problem in computer vision. We adopt the LEMT method [3, 4] to extract a long-term common shape and a set of relative denser long-term motion trajectories from each video sequence. Then we present a hybrid feature representation to integrate the extracted shape and motion features. Furthermore, we learn a RMM for each action category to be utilized for action classification. The RMMs are more informative than pure BOF methods by considering the spatial layout of features and are able to overcome the difficulty of body parts segmentation in LR videos. Experimental results show the effectiveness of our approach. As future work, we would like to develop an algorithm to automatically adjust the number of regions in each RMM to get the best accuracy.

7. REFERENCES

[1] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[2] Chia-Chih Chen and JK Aggarwal, "Recognizing human action from a far field of view," in *IEEE Workshop on Motion and Video Computing*, 2009.

[3] Huijun Di, Linmi Tao, and Guangyou Xu, "A mixture of transformed hidden markov models for elastic motion estimation," *IEEE TPAMI*, vol. 31, no. 10, pp. 1817–1830, 2009.

[4] Lv Feng, Di Hui-Jun, Lu Yao, and Xu Guang-You, "Non-rigid tracking method based on layered elastic motion analysis," *Acta Automatica Sinica*, vol. 41, no. 2, pp. 295–303, 2015.

[5] Alexei Efros, Alexander C Berg, Greg Mori, Jitendra Malik, et al., "Recognizing action at a distance," in *IEEE ICCV*, 2003.

[6] MS Ryoo, Chia-Chih Chen, JK Aggarwal, and Amit Roy-Chowdhury, "An overview of contest on semantic description of human activities (SDHA) 2010," in *Recognizing Patterns in Signals, Speech, Images and Videos*, pp. 270–285. Springer, 2010.

[7] Kai Guo, Prakash Ishwar, and Janusz Konrad, "Action recognition in video by sparse representation on covariance manifolds of silhouette tunnels," in *Recognizing Patterns in Signals, Speech, Images and Videos*, pp. 294–305. Springer, 2010.

[8] Arridhana Ciptadi, Matthew S Goodwin, and James M Rehg, "Movement pattern histogram for action recognition and retrieval," in *ECCV*, 2014.

[9] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *IEEE TPAMI*, vol. 29, no. 12, pp. 2247–2253, 2007.

[10] Hejin Yuan, "A semi-supervised human action recognition algorithm based on skeleton feature," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 1, pp. 175–182, 2015.

[11] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *PRL*, vol. 34, no. 15, pp. 1799–1807, 2013.

[12] Kai Guo, Prakash Ishwar, and Janusz Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[13] John Canny, "A computational approach to edge detection," *IEEE TPAMI*, vol. 8, no. 6, pp. 679–698, 1986.

[14] Chia-Chih Chen, M.S.Ryoo, and J.K.Aggarwal, "UT-Tower dataset: aerial view activity classification challenge," http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html, 2010.

[15] Yang Yang, Imran Saleemi, and Mubarak Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE TPAMI*, vol. 35, no. 7, pp. 1635–1648, 2013.

[16] Kishore K Reddy, Naresh Cuntoor, Amitha Perera, and Anthony Hoogs, "Human action recognition in large-scale datasets using histogram of spatiotemporal gradients," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2012.

[17] Chia-Chih Chen and Jake K Aggarwal, "Human shadow removal with unknown light source," in *IEEE ICPR*, 2010, pp. 2407–2410.

[18] Kun Zhao, Arnold Wiliem, and Brian Lovell, "Kernelised orthonormal random projection on grassmann manifolds with applications to action and gait-based gender recognition," in *IEEE International Conference on Identity, Security and Behavior Analysis*, 2015.

[19] Xianbin Cao, Bo Ning, Pingkun Yan, and Xuelong Li, "Selecting key poses on manifold for pairwise action recognition," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 168–177, 2012.

[20] Snehasis Mukherjee, Sujoy Kumar Biswas, and Dipti Prasad Mukherjee, "Recognizing human action at a distance in video by key poses," *IEEE T-CSVT*, vol. 21, no. 9, pp. 1228–1241, 2011.