Leveraging Textural Features for Recognizing Actions in Low Quality Videos

Saimunur Rahman[1], John See[2], and Chiung Ching Ho[3]

Centre of Visual Computing, Faculty of Computing and Informatics
Multimedia University, Cyberjaya 63100, Selangor, Malaysia
[1]saimunur.rahman14@student.mmu.edu.my
{[2]johnsee,[3]ccho}@mmu.edu.my

Abstract. Human action recognition is a well researched problem, which is considerably more challenging when video quality is poor. In this paper, we investigate human action recognition in low quality videos by leveraging the robustness of textural features to better characterize actions, instead of relying on shape and motion features may fail under noisy conditions. To accommodate videos, texture descriptors are extended to three orthogonal planes (TOP) to extract spatio-temporal features. Extensive experiments were conducted on low quality versions of the KTH and HMDB51 datasets to evaluate the performance of our proposed approaches against standard baselines. Experimental results and further analysis demonstrated the usefulness of textural features in improving the capability of recognizing human actions from low quality videos.

1 Introduction

Recognizing human actions [1–9] from unconstrained videos is of central importance in a variety of real-world applications such as video surveillance, human-computer interaction and video retrieval & indexing . Actions in video present a wide range of variations, ranging from object-based variations such as appearance, view pose and occlusion, to more challenging scene-related problems such as illumination changes, shadows, and camera motions. One relatively understudied problem is video quality [8, 10, 11]. Current video processing research have focused primarily on good quality videos that offer fine details and strong signal fidelity which are not feasible for real-time video processing, or lightweight mobile applications. Among recently proposed methods, local spatio-temporal handcrafted features such as space-time interest points (STIPs) [12], cuboids [6], extended SURF [6], dense sampling [6] and dense trajectories [7] are popular for their simplicity and effectiveness in human action recognition. A majority of these methods [2, 6, 7] used HOG and HOF descriptors to characterize shape and motion information. Meanwhile, spatio-temporal textural features such as LBP-TOP [13] and extended LBP-TOP [5] have also found its way to action recognition, albeit in less celebrated fashion. Their reported performances were promising, though they acknowledge the lack the explicit encoding of motion features. The use of textures is less common in literature, though there are promising benefits that can be established in our recent works [10, 11]. Following the emphasis on understanding the behavior of existing approaches over video quality [14], our recent works [10, 11] evaluated STIPs [12] on spatially and temporally downsampled videos and showed that they are not effective with the deterioration of video quality. It was also shown that, the individual shortcomings of shape and motion features [2] can be alleviated by using complementary textural features [13].

Motivated by the analysis above, we aim to investigate and present viable approaches to the problem of human action recognition in low-quality video. In this paper, we propose approaches that utilize textural features in addition to conventional space-time shape and motion features to vastly improve the recognition of human actions under such conditions. We conduct an extensive series of experiments on poor quality versions/subsets of two publicly available action benchmark datasets: The classic KTH [1] and the large-scale HMDB51 [8]. The rest of the paper is organized as follows: Section 2 describes the feature descriptors used, Section 3 elaborates on the experiments conducted with its results and further analysis. Finally, Section 4 concludes the paper.

2 Spatio-Temporal Feature Representation

This section presents the descriptors used for extracting features from each video. Section 3.1 describes shape and motion feature descriptors while Section 3.2 describes various textural feature descriptors employed in this work.

2.1 Shape and Motion Features

For extraction of shape and motion features from video, the Harris3D [12] detector (a space-time extension of the popular Harris detector) was used as the local spatio-temporal interest point (STIP) detector. It detects local structures where image values have significant local variations in both space and time. To characterize the shape and motion information accumulated in space-time neighborhoods of the detected STIPs, we applied Histogram of Gradient (HOG) and Histogram of Optical Flow (HOF) feature descriptors as proposed in [12]. The combination of HOG/HOF descriptors produces descriptors of size $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma, \Delta_t(\tau) = 8\tau$ (σ and τ are the spatial and temporal scales). Each volume is subdivided into a $n_x \times n_y \times n_t$ grid of cells; for each cell, 4-bin histograms of gradient orientations (HOG) and 5-bin histograms of optical flow (HOF) are computed [6]. We used the original implementation and standard parameter settings $n_x, n_y = 3, n_t = 2$ defined in [6].

2.2 Textural Features

For the extraction of textural features we employed three feature descriptors namely LBP, LPQ and BSIF, which are then extended by three orthogonal planes (TOP). They are briefly discussed as follows:

LBP features: Local binary patterns (LBP) [15] are used to describe the structural variations of brightness (contrast) in an image. The LBP operator uses center pixel as threshold and label its circular neighborhood within radii R by 1 if larger than center, otherwise label by 0 if smaller than center. The binary code for each center pixel is formed by multiplying binalized values obtained by thresholding with corresponding (given) pixel weights and summing them up. The $LBP_{P,R}$ operator produces 2^P different output values, corresponding to the 2^P different binary patterns that can be formed by the P pixels in the neighborhood set.

LPQ features: Local phase quantization [16] operator uses local phase informations to produce blur-invariant image features extracted by computing short term Fourier transform (STFT) in rectangular neighborhoods N_x, which are defined as:

$$F(u,x) = \sum_{y \in N_x} f(x-y) e^{-j2\rho u^T y} = \mathbf{W}_u^T \mathbf{f}_x$$

where, W_u is the basis vector and f_x is image samples across N_x. Four complex coefficients corresponds to 2D frequencies is considered for forming LPQ features: $u_1 = [a, 0]^T$, $u_1 = [0, a]^T$, $u_1 = [a, a]^T$ and $u_1 = [a, -a]^T$, where a is a scalar. To express the phase informations, a binary coefficient is then formed from the sign of imaginary and real part of these Fourier coefficients. An image is then produced by representing 8 binary values (obtained from binary coefficient) as the integer value between 0 to 255. Finally, LPQ feature histogram is then constructed from the produced image.

BSIF features: Binarized statistical image features (BSIF) [17] is a recently proposed method that efficiently encodes texture information, in a similar vein to earlier methods that produce binary codes [15, 16]. Given an image X of size $p \times p$, BSIF applies a linear filter F_i learnt from natural images through independent component analysis (ICA), on the pixel values of X and obtained the filter response,

$$r_i = \sum_{u,v} F_i(u,v) X(u,v) = \mathbf{f}_i^T \mathbf{x}$$

where $\mathbf{f}$ and $\mathbf{x}$ are the vectorized form of F_i and W respectively. The binarized feature b_i is then obtained by thresholding r_i at the level zero, i.e. $b_i = 1$ if $r_i > 0$ and $b_i = 0$ otherwise. The decomposition of the filter mask F_i allows the independent components or basis vectors to be learnt by ICA. Succinctly, we can learn n number of $l \times l$ linear filters W_i, stacked into a matrix $\mathbf{W}$ such that all responses can be efficiently computed by $\mathbf{s} = \mathbf{Wx}$. Consequently, an n-bit binary code is produced for each pixel, which then builds the feature histogram for the image.

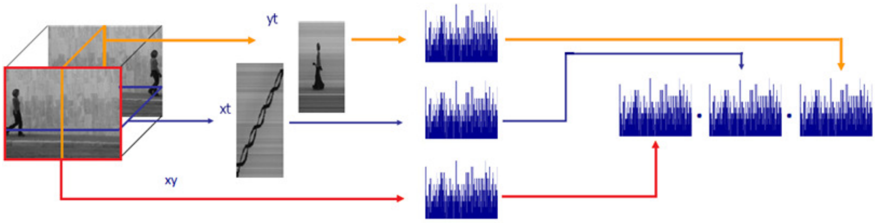

Fig. 1. Three Orthogonal Planes (TOP) approach for describing spatio-temporal textures. The histograms from the XY, XT and YT planes are concatenated to construct the final histogram.

Spatio-temporal extension of textural features: Inspired by the success of works relating to the recognition of dynamic sequences based on the LBP [5, 13], we consider the three orthogonal planes (TOP) approach which extends the aforementioned texture features to extract features for our task. An illustration of the TOP extension is shown in Figure 1. Given a volumetric space of $X \times Y \times T$, the TOP approach extracts the texture descriptors along the XY, XT and YT orthogonal planes where, the XY plane encodes structural information while XT and YT planes encode space-time transitional information. The histograms of all three planes are concatenated to form the final feature histogram.

In this work, we have applied parameter settings of $LBP-TOP_{8,8,8,2,2,2}$ with non-uniform patterns as specified in [5, 10], $LPQ-TOP_{5,5,5}$ as specified in [18] and 9x9 12 bit ($l = 9$, $n = 12$) filters for BSIF-TOP.

3 Experiments

In this section, we describe the datasets used in the experiments and report the results on various approaches. We also compare the effectiveness of different textural features, and discuss their computational costs.

3.1 Datasets

We conduct our experiments on two public datasets: KTH [1] and HMDB [8], in a manner that is suitable for our work. The **KTH** action dataset is one of the most popular datasets for action recognition, consisting of videos captured from a rather controlled environment. It contains 6 action classes performed by 25 actors in 4 different scenarios. There are 599 video samples in total (one subject has one less clip) and each clip is sampled at 25 *fps* at a frame resolution of 160 × 120 pixels. We follow the original experimental setup specified in [1], reporting the average accuracy over all classes. Similar to our previous work [10, 11], six downsampled versions of the KTH were created – three for spatial downsampling (SD_α), and three for temporal downsampling (TD_β). We limit our experiments to downsampling factors, $\alpha, \beta = \{2, 3, 4\}$, which denotes spatial

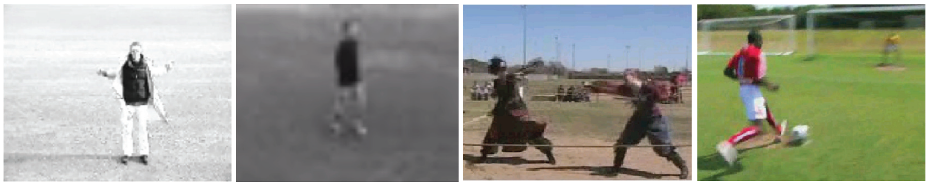

Fig. 2. Sample video frames from *(left two)* KTH (downsampled) and *(right two)* HMDB ('bad' and 'medium' clips) datasets.

or temporal downsampled versions of half, a third and a fourth of the original resolution or frame rate respectively.

The **HMDB** is one of the largest human action recognition dataset that is fast gaining popularity. It has a total of 6,766 videos of 51 action classes collected from movies or YouTube videos. HMDB is a considerably challenging dataset with videos acquired from uncontrolled environment with large viewpoint, scale, background, illumination variations. Videos in HMDB are annotated with a rich set of meta-labels including quality information; three quality labels were used, i.e. 'good', 'medium' and 'bad'. Three training-testing splits were defined for the purpose of evaluations, and performance is to be reported by the average accuracy over all three splits. In our experiments, we use the same specified splits for training, while testing was done using only videos with 'bad' and 'medium' labels; for clarity, they are respectively indicated as **HMDB-BQ** and **HMDB-MQ** hereafter. In the 'medium' quality videos, only large body parts are identifiable, while they are totally unidentifiable in the 'bad' quality videos due to the presence of motion blur and compression artifacts. 'Bad' and 'medium' videos comprise of 20.8% and 62.1% of the total number of videos, respectively. Figure 2 shows some sample frames of various actions from the downsampled KTH and poor quality HMDB subsets.

3.2 Evaluation Framework

We evaluated our methods using traditional bag-of-visual-words representation where, visual features are represented as histogram of visual codewords obtained from hard assignment by vector quantization (VQ). Classification is performed with a non-linear multi-class support vector machine (SVM) with χ^2-kernel, adopting a one-versus-all strategy. We use a computationally efficient approximation of the non-linear kernel by Vedaldi et al. [19] which allows features to undergo a non-linear kernel map expansion before a linear SVM classification. This also provides us the flexibility of deciding which features are to be "kernelized". We follow the settings specified in [6] which are shown to be effective across various datasets, i.e. $k = 4000$ and ℓ_2-normalization.

3.3 Experimental Results and Analysis

The experimental results are summarized in Tables 1 and 2. In our experiments, we chose to concatenate the quantized HOG and HOF descriptors of the STIPs

Table 1. Comparison of different texture feature combinations on various downsampled versions of KTH

Method	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
HOG/HOF [6]	83.33	76.39	65.74	86.11	81.94	76.85
HOG+HOF [10]	84.26	80.09	75.46	87.04	80.09	81.48
HOG+HOF + LBP-TOP [11]	87.41	80.74	77.69	87.87	82.50	80.37
HOG+HOF + LPQ-TOP	88.15	81.30	78.52	87.50	81.85	80.00
HOG+HOF + BSIF-TOP	**89.07**	**85.00**	**81.67**	**88.52**	**87.04**	**84.91**

Table 2. Comparison of various texture feature combinations HMDB 'bad' (HMDB-BQ) and 'medium' (HMDB-MQ) subsets

Method	HMDB-BQ	HMDB-MQ
HOG/HOF [8]	17.18	18.68
C2 [8]	17.54	23.10
HOG+HOF [10]	21.71	23.68
HOG+HOF + LBP-TOP [11]	20.80	24.20
HOG+HOF + LPQ-TOP	23.89	28.36
HOG+HOF + BSIF-TOP	**32.46**	**37.14**

('histogram-level' concatenation), as denoted by "HOG+HOF". This representation is found to be generally more effective than a 'descriptor-level' concatenation (denoted by "HOG/HOF") which was originally used in [2, 6]. Meanwhile, textural features i.e. LBP-TOP, LPQ-TOP, BSIF-TOP, are extracted from the entire video volume as feature histograms, and then aggregated with the HOG and HOF histograms.

Overall, all three approaches that utilize the additional textural features clearly demonstrate significant improvement, as compared to the baseline methods. This is consistent across both the downsampled KTH data (Table 1) and HMDB poor quality subsets (Table 2). Among the evaluated spatio-temporal textural features, the BSIF-TOP appears to be the most promising choice, as it outperforms the other approaches.

Figure 3 offers a closer look at how the BSIF-TOP fare against the other two features across varying downsampled (spatially and temporally) versions of the KTH dataset. Evidently, BSIF-TOP performs distinctly better than the LBP-TOP and LPQ-TOP features as the spatial resolution and temporal sampling rate drops. It is about 4% better than the LBP-TOP on the spatially downsampled data, and about 5% better than the LPQ-TOP on the temporally downsampled data. On the HMDB subsets, the approaches that incorporated textural features are clearly better than its original HOG/HOF and C2 baseline methods [8]. By aggregating BSIF-TOP textures, recognition capability of both HMDB-BQ and HMDB-MQ improves to almost double that of the original baseline results.

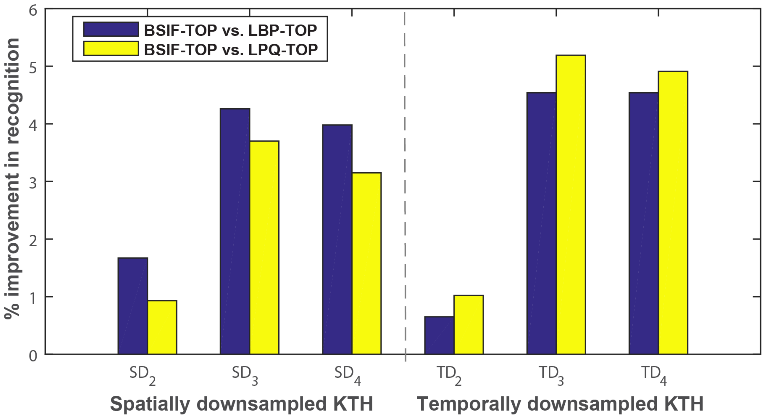

Fig. 3. Percentage improvement of BSIF-TOP over LBP-TOP and LPQ-TOP, when aggregated with STIP (HOG+HOF) features

Table 3. Computational cost of different feature descriptors

	HOG+HOF	LBP-TOP	LPQ-TOP	BSIF-TOP
Time (in sec.)	13.76	47.57	2.48	5.25

3.4 Computational Complexity

Using a Core i7 32GB RAM machine, we compare the speed (includes feature detection and quantization time) of computing different feature descriptors, as shown in Table 3.4. This computational test was performed on a sample video from SD_2 version of KTH dataset consist of 656 image frames. Among the textural features, the LPQ-TOP and BSIF-TOP are the most efficient methods (both much quicker than computing HOG+HOF on STIPs), and yet they are able to contribute significantly to the recognition accuracy.

4 Conclusion

Shape, motion and textural features are all important features for recognizing human actions. In this paper, we leveraged on textural features to improve the recognition of human actions in low quality video clips. Considering that most current approaches involved only shape and motion features, the use of spatio-temporal textural features is a novel proposition that improves the recognition performance by a good margin. Among all, the usage of BSIF-TOP dynamic textures is most promising, with a significant leap of around 16% and 18% on the KTH-SD_4 and HMDB-MQ datasets respectively, over their original baselines. In future, we intend to extend this work towards a larger variety of human action datasets. It is also worth designing textural features that are more discriminative and robust towards complex backgrounds.

References

1. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proc. of Int. Conf. on Pattern Recognition. (2004) 32–36
2. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE CVPR. (2008) 1–8
3. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008) 275–1
4. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. In: BMVC. Volume 1. (2008) 2
5. Mattivi, R., Shao, L.: Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In: Proc. of CAIP. (2009) 740–747
6. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009) 124–1
7. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: IEEE CVPR. (2011) 3169–3176
8. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: IEEE ICCV. (2011) 2556–2563
9. Wang, X., Wang, L., Qiao, Y.: A comparative study of encoding, pooling and normalization methods for action recognition. In: Proc. of ACCV. (2013) 572–585
10. Rahman, S., See, J., Ho, C.C.: Action recognition in low quality videos by jointly using shape, motion and texture features. In: IEEE Int. Conf. on Signal and Image Processing App. (2015) *To appear*
11. See, J., Rahman, S.: On the effects of low video quality in human action recognition. In: Digital Image Computing: Techniques and Applications (DICTA). (2015) *To appear*
12. Laptev, I.: On space-time interest points. Int. Journal of Computer Vision **64**(2-3) (2005) 107–123
13. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE PAMI **29**(6) (2007) 915–928
14. Oh, S., Hoogs, A., Perera, A., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: IEEE CVPR. (2011) 3153–3160
15. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Trans. PAMI **28**(12) (2006) 2037–2041
16. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE PAMI **24**(7) (2002) 971–987
17. Kannala, J., Rahtu, E.: Bsif: Binarized statistical image features. In: Pattern Recognition (ICPR), 2012 21st Int. Conf. on. (2012) 1363–1366
18. Päivärinta, J., Rahtu, E., Heikkilä, J.: Volume local phase quantization for blur-insensitive dynamic texture classification. In: Image Analysis. Springer (2011) 360–369
19. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE PAMI **34**(3) (2012) 480–492