SPATIO-TEMPORAL MID-LEVEL FEATURE BANK FOR ACTION RECOGNITION IN LOW QUALITY VIDEO

Saimunur Rahman, John See

Centre of Visual Computing, Faculty of Computing and Informatics
Multimedia University
Cyberjaya 63100, Malaysia

ABSTRACT

It is a great challenge to perform high level recognition tasks on videos that are poor in quality. In this paper, we propose a new spatio-temporal mid-level (STEM) feature bank for recognizing human actions in low quality videos. The feature bank comprises of a trio of local spatio-temporal features, i.e. shape, motion and textures, which respectively encode structural, dynamic and statistical information in video. These features are encoded into mid-level representations and aggregated to construct STEM. Based on the recent binarized statistical image feature (BSIF), we also design a new spatio-temporal textural feature that extracts discriminately from 3D salient patches. Extensive experiments on the poor quality versions/subsets of the KTH and HMDB51 datasets demonstrate the effectiveness of the proposed approach.

***Index Terms*—** Action recognition, Low quality video, Mid-level representation, Texture features, BSIF

1. INTRODUCTION

Action recognition [1, 2, 3, 4, 5] is becoming increasingly important today due to its various application domains such as video surveillance, video indexing and searching, and human-computer interaction. However, action recognition in real world scenarios still remains a challenging issue especially concerning video quality [6, 7, 8]; typical problems include low resolution and frame rates, compression artifacts, background clutter, camera ego-motion and jitter. Despite advances in video technology, there is still an undeniable need for efficient processing, storage and transmission. As such, it is crucial to deal with the problem of low video quality by designing more robust approaches to action recognition.

In recent years, various methods have been developed to recognize human actions from video. Currently, among handcrafted methods, *shape* and *motion* are the most widely used by the action recognition community. The extraction of these features consists of two essential steps – a *detection* step where important points or salient regions are extracted from the video, and *description* step which then describes the patterns from the extracted region. Among typical detectors include space-time interest points [9], cuboids [10], dense sampling [1] and dense trajectories [11]. The HOG and HOF features [1, 9] appeared most prominently in recent state-of-the-art approaches owing to their effectiveness in characterizing dynamic and structural properties of actions. However, their reliance on localized feature regions may render them ineffective when discriminating between actions in low video quality [7].

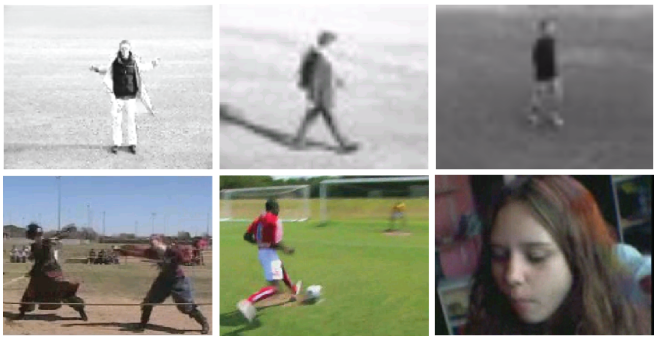

Fig. 1. Low quality videos generally affected by poor resolution, sampling rate, motion blur and compression artifacts. Here are sample video frames from *(top row)* KTH (downsampled version) and *(bottom row)* HMDB51 (with 'bad' quality label) datasets.

The use of *textural* features are less common in action recognition; among proposed representations include LBP-TOP [12] and Extended LBP-TOP [13]. These methods use the notion of three orthogonal planes (TOP) to extend static image-based textures to spatio-temporal dynamic textures. More recently, Kannala & Rahtu proposed binarized statistical image features (BSIF) [14] which have showed tremendous potential compared to its predecessors. While these methods were able to show effective result across different action datasets, our recent work [7] has shown that local STIP features [9] become increasingly ineffective when video quality deteriorate spatially and temporally. It was shown that this problem can be alleviated by introducing complementary robust global textural features. Moreover, the holistic nature of

these descriptors often produce indiscriminate features since illumination variations or unrelated motion from videos with complex background can be erroneously regarded as textures.

Motivated by the analyses above, this paper proposes a new spatio-temporal mid-level (STEM) feature bank for recognizing actions in low quality videos. STEM is designed to combine the benefits of local explicit patterns surrounding interest points, and global statistical patterns. Specifically, we first detect features at the local and global levels, or *streams*; for each stream, we use state-of-the-art spatio-temporal approaches [9, 14] to describe its respective patterns. More significantly, our textural features are compactly extracted from 3D salient patches. To build the STEM feature bank consisting of a trio of features, we integrate shape-motion descriptors soft-quantized by Fisher Kernel encoding, and the discriminative textural descriptors. Finally, we show the efficacy of the proposed STEM feature bank in poor quality versions/subsets of two public datasets – KTH and HMDB51.

2. PROPOSED FEATURE BANK

A graphical overview of our proposed feature bank is illustrated in Figure 2. An input video undergoes series of steps, in two separate streams: *local* and *global*. In the local stream, spatio-temporal interest points are first detected and represented with shape-motion descriptors. The global stream involves the extraction of spatio-temporal textural features, which are discriminately selected based on visual saliency. To obtain a compact mid-level representation, the local shape-motion features are encoded by Fisher Vectors (FV) while the global textural features are built based on regions defined by the 3D saliency mask. The concatenation of both sets produces the spatio-temporal mid-level (STEM) feature bank, which is subsequently used for recognition. We present the shape-motion features and salient textural features in sections 2.1 and 2.2 respectively.

2.1. Shape-Motion Features

To extract shape-motion features from video, we employ Harris3D [9] detector (a space-time extension of the popular Harris detector) 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 describe 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). 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 video volume is subdivided into $n_x \times n_y \times n_t$ grid of cells; for each cell, 4-bin histograms of gradient orientations (HOG) and 5-bin histograms of optical flow (HOF) are computed [1]. In our work, we opt for grid parameters $n_x, n_y = 3, n_t = 2$ for all videos, as suggested in the original paper. Finally, these descriptors are further encoded into Fisher Vectors (FV) [15], whereby the descriptors undergo soft quantization by fitting Gaussian Mixture Models (GMM) to the distribution of descriptors.

2.2. Salient Textural Features

This section describes a new video based textural representation, called salient textures. We first present the texture detection method we used. Then, we will describe our spatio-temporal textural feature representation technique. Finally, based on the saliency detection, we discuss the details of salient texture calculation.

Textural feature detection: Binarized statistical image features (BSIF) [14] is a recently proposed method that efficiently encodes texture information, in a similar vein to earlier methods that produce binary codes [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,

$$r_i = \sum_{u,v} F_i(u,v)X(u,v) = \mathbf{f}_i^T \mathbf{x} \tag{1}$$

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 $\mathbf{s} = \mathbf{W}\mathbf{x}$. Consequently, an n-bit binary code is produced for each pixel, which then builds the feature histogram for the image.

Spatio-temporal textural features: Inspired by the recent success of [17, 18] in various dynamic recognition tasks, we consider the three orthogonal planes (TOP) approach to extend the BSIF feature to extract spatio-temporal dynamic textures. Given a volumetric space of $X \times Y \times T$, the BSIF histogram can be defined as

$$h_j^{plane} = \sum_{p \in plane} \mathcal{I}\{b_i(p) = j\} \tag{2}$$

where $j \in \{1, \ldots, 2^n\}$, p is a pixel at a location (x, y, t) in a specific plane, and $\mathcal{I}\{.\}$ is a function indicating 1 if true, and 0 otherwise. The histogram bins of each plane are then normalized to get a coherent description, $\tilde{\mathbf{h}}^{plane} = \{\tilde{h}_1^{plane}, \ldots, \tilde{h}_{2^n}^{plane}\}$. Finally, we concatenate the histograms of all three planes,

$$\mathbf{H} = \{\tilde{\mathbf{h}}^{XY}, \tilde{\mathbf{h}}^{XT}, \tilde{\mathbf{h}}^{YT}\} \tag{3}$$

to form the BSIF-TOP textural feature (of length $3 \cdot 2^n$) In this work, we apply 12, 9x9 filters ($n = 12$), which was chosen empirically, across all video samples.

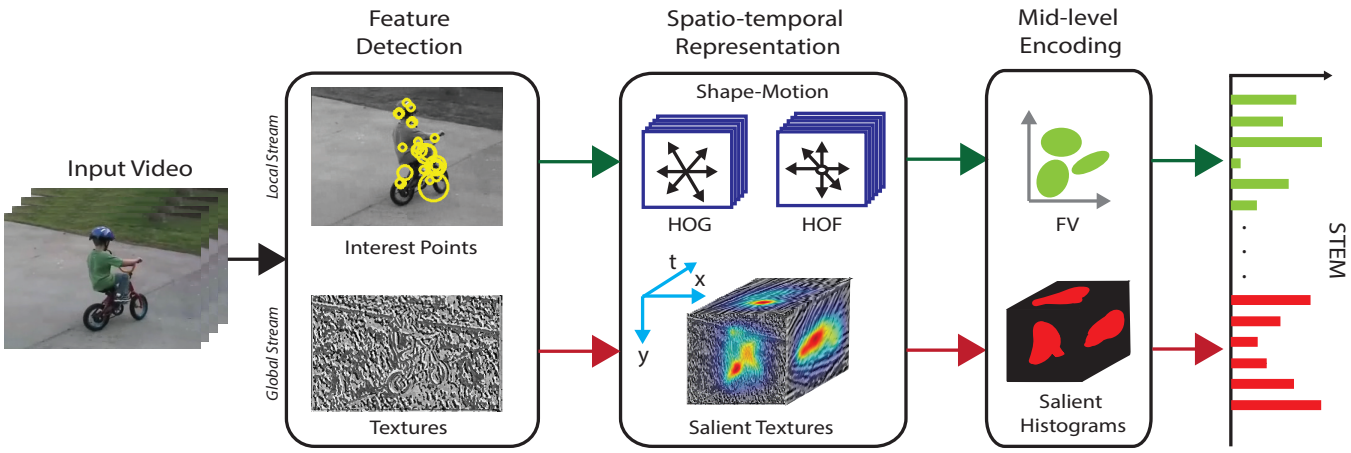

Fig. 2. Illustration of the proposed STEM feature bank

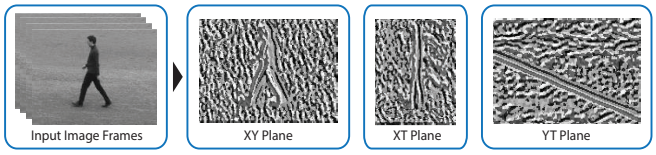

Fig. 3. A sample image frame and its BSIF code images in the XY, XT and YT planes

Salient texture formation: In an action video such as running or fencing, the runner or fencers gets the most visual attention across the video, hence they comprise of the most salient locations in video. As such, we apply saliency to the textural features extracted from each video. Motion- or video-based saliency methods are not always effective in constructing equitable saliency maps in presence of large camera motion and noisy flow fields [2]. In our work, we adapt graph-based visual saliency [19] to capture the salient regions in each frame. This method is computationally simple and is able to model natural fixation based on a variety of features.

Firstly, we calculate a set of three feature maps: A *contrast* map computed using luminance variance in a local neighborhood of size one-tenth of the image width, four *orientation* maps computed using Gabor filters (at orientations $0°$, $45°$, $90°$, $135°$), and a *flicker* map by simple frame differencing. Next, a fully connected directed graph $\mathcal{G}$ is constructed to extract activation maps, where the weight of edge between two nodes corresponding to pixels at location(x, y) and (m, n) can be expressed as:

$$w^{act}_{x,y,m,n} = |M_{x,y} - M_{m,n}|\, e^{\left(-\frac{(x-m)^2+(y-n)^2}{2\sigma_1^2}\right)} \quad (4)$$

where σ_1 is a free parameter. The calculated graph with normalized weights is used to construct a Markovian chain [19] where achieving equilibrium distribution results in an activation map $A_{m,n}$. To normalize the activation map, another graph is constructed with each coordinate from $A_{m,n}$ representing the nodes, and the weights are defined as,

$$w^{norm}_{x,y,m,n} = A_{m,n}\, e^{\left(-\frac{(x-m)^2+(y-n)^2}{2\sigma_2^2}\right)} \quad (5)$$

When equilibrium distribution is achieved, mass will be concentrated at nodes with high activation, thus forming the final saliency map, $S_{x,y}$

Finally, for each frame we convert the saliency map $S_{x,y}$ into a binary saliency mask $Z_{x,y}$ by utilizing Otsu's method [20]. We choose this method because of its capability of locating an optimal threshold that optimizes the foreground and background pixel distributions. This yields a 3D salient mask, $Z(p)$ with p a pixel at location (x, y, t). Applying saliency to Eq. 2, computation of the j-th bin for the new salient BSIF histogram is redefined as

$$\bar{h}^{plane}_j = \sum_{p \in plane} \mathcal{I}\{\{b_i(p) = j\} \cap \{Z(p) = 1\}\} \quad (6)$$

3. EXPERIMENTS

In this section, we first describe the datasets used and their respective evaluation setup and protocol. Then, we report experimental results and further discussions and analysis.

3.1. Datasets

We conduct our experiments on two popular publicly available datasets: KTH and HMDB51, in a manner which serves the purpose of our work. **KTH** has 599 videos in total (one class contains one video less) and each clip is sampled at 25 fps at frame resolution 160×120 pixels, lasting between 10-15 seconds. Following our previous work [7], six different downsampled versions of the KTH are created –

Table 1. Recognition accuracy (%) of various feature approaches on the KTH and HMDB51 low quality versions/subsets

Method	KTH						HMDB	
	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4	BQ	MQ
HOG+HOF (BoW encoding) [1, 4]	88.24	81.11	73.89	87.04	82.87	82.41	17.40	22.77
HOG+HOF [5]	89.63	82.31	78.98	89.35	86.11	83.89	26.02	30.53
HOG+HOF+LBP-TOP [8]	89.81	81.48	78.70	89.35	86.11	84.72	28.49	35.24
STEM (w/o salient textures)	89.35	82.87	79.72	89.63	87.41	84.63	33.78	38.76
STEM	88.52	**83.98**	**83.15**	**90.00**	**88.06**	**85.09**	**34.08**	**38.94**

three for spatial downsampling (SD_α), and three for temporal downsampling (TD_β) where, the downsampling factors $\alpha, \beta = \{2, 3, 4\}$ denote the factor in which the original video is downsampled (e.g. TD_3 denotes reducing the original frame rate by a third). We follow the original protocol specified in [3], by reporting the average accuracy over all classes.

HMDB51 [4] contains 6,766 videos across 51 action classes and is one of the largest action dataset available recently. All the videos in HMDB51 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 specified for the purpose of evaluation, 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. 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. For clarity, we specify **HMDB51-MQ** and **HMDB51-BQ**, to denote the 'medium' and 'bad' quality subsets of this dataset.

3.2. Evaluation Setup

We now define other evaluation parameters and methods used. Following [5, 15], we used $k = 256$ for FV encoding, applying power and ℓ_2–normalization. The free parameters in Eq. 4 and 5 (σ_1, σ_2) are fixed to 0.15 and 0.06 of the map width respectively, following the author's implementation [19]. For classification, we determine the action labels by employing a multi-class support vector machine (SVM) with homogeneous χ^2-kernel [21].

3.3. Experimental Results

This section presents the experimental results of our proposed STEM feature bank on the six KTH downsampled versions and two HMDB51 subsets. Table 1 shows all the results. We also provide comparisons with other recent approaches.

Results on KTH dataset: We chose the KTH to perform this extensive downsampled experiment as it is lightweight and a widely used benchmark among most public datasets. We can observe that as the video quality deteriorates (particularly the spatial resolution), most methods struggle to maintain their original performances (i.e. 92.13% for [5], 91.8% for [1]). From the results in Table 1, the proposed STEM outperforms other feature approaches across all modes except in SD_2. This shows that STEM features are more robust towards poorer video quality; the improvement most obvious when dealing with spatial resolution. For a better perspective, the STEM approach garnered 93.43% on the original KTH data, which is only marginally better than the other methods. We observe that the discriminative nature of the salient textures in STEM plays a significant role in obtaining better performance compared to its non-salient counterpart, especially in SD_4 and TD_4 videos.

Results on HMDB51 subsets: We also tested on the HMDB51 low quality subsets in order to evaluate the effectiveness of our proposed method on larger number of classes. Results in Table 1 show the superiority of the STEM feature bank over other feature approaches. The salient textural features which are produced on the global stream clearly helps improve the recognition accuracy by $\approx$ 8% on both BQ ('bad') and MQ ('medium') subsets compared to spatio-temporal interest point descriptors. Interestingly, the use of salient textures in STEM could only marginally surpass that of non-salient textures. This is likely due to the complexity of background scenes in this dataset.

Multi-scale salient texture: To further enrich the statistical information encoded in our salient texture descriptor we extended our texture formation to a multi-scale variety by employing a number of filters of different sizes, specifically $l = \{3, 9, 15\}$. This is able to increase the accuracy of STEM on both datasets by $\sim$ 1-2% but at the expense of higher computational load. This direction can be further explored.

4. CONCLUSION

In this work, we have presented a new spatio-temporal mid-level (STEM) feature bank that integrates the advantages of local explicit patterns from interest points and global salient statistical patches. In comparison to state-of-the-art, our proposed method achieved superior recognition performance on low quality versions or subsets of two public datasets. In future, we plan to extend our feature bank to accommodate

denser types of features while also increasing the robustness of the saliency method used.

5. REFERENCES

[1] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*. BMVA Press, 2009, pp. 124–1.

[2] Waqas Sultani and Imran Saleemi, "Human action recognition across datasets by foreground-weighted histogram decomposition," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 764–771.

[3] Christian Schüldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Proc. of Int. Conf. on Pattern Recognition*, 2004, pp. 32–36.

[4] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "HMDB: A large video database for human motion recognition," in *IEEE ICCV*, 2011, pp. 2556–2563.

[5] Xingxing Wang, LiMin Wang, and Yu Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," in *Proc. of ACCV*, pp. 572–585. 2013.

[6] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al., "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3153–3160.

[7] Saimunur Rahman, John See, and Chiung Ching Ho, "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, p. To appear.

[8] John See and Saimunur Rahman, "On the effects of low video quality in human action recognition," in *Digital Image Computing: Techniques and Applications (DICTA)*, 2015, p. To appear.

[9] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[10] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[11] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[12] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen, "Human activity recognition using a dynamic texture based method.," in *BMVC*, 2008, vol. 1, p. 2.

[13] Riccardo Mattivi and Ling Shao, "Human action recognition using lbp-top as sparse spatio-temporal feature descriptor," in *Computer Analysis of Images and Patterns*. Springer, 2009, pp. 740–747.

[14] Juho Kannala and Esa Rahtu, "Bsif: Binarized statistical image features," in *ICPR*. IEEE, 2012, pp. 1363–1366.

[15] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. of ECCV*, pp. 143–156. 2010.

[16] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam, "Survey on lbp based texture descriptors for image classification," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3634–3641, 2012.

[17] Guoying Zhao and Matti Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE PAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[18] Juhani Päivärinta, Esa Rahtu, and Janne Heikkilä, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Image Analysis*, pp. 360–369. Springer, 2011.

[19] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[20] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[21] Andrea Vedaldi and Andrew Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE PAMI*, vol. 34, no. 3, pp. 480–492, 2012.