2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)

Action Recognition in Low Quality Videos by Jointly Using Shape, Motion and Texture Features

Saimunur Rahman*, John See*†, Chiung Ching Ho *‡
*Centre of Visual Computing, Faculty of Computing and Informatics
Multimedia University
Cyberjaya 63100, Selangor, Malaysia
Email: saimunur.rahman14@student.mmu.edu.my
† Email: johnsee@mmu.edu.my
‡ Email: ccho@mmu.edu.my

***Abstract*—Shape, motion and texture features have recently gained much popularity in their use for human action recognition. While many of these descriptors have been shown to work well against challenging variations such as appearance, pose and illumination, the problem of low video quality is relatively unexplored. In this paper, we propose a new idea of jointly employing these three features within a standard bag-of-features framework to recognize actions in low quality videos. The performance of these features were extensively evaluated and analyzed under three spatial downsampling and three temporal downsampling modes. Experiments conducted on the KTH and Weizmann datasets with several combination of features and settings showed the importance of all three features (HOG, HOF, LBP-TOP), and how low quality videos can benefit from the robustness of textural features.**

I. Introduction

Human action recognition in video is an active area of research in computer vision, with many applications in various fields including video surveillance, content-based video archiving and browsing, and human computer interaction.

Actions in video undergo a wide range of variations such as size, appearance and view pose, while more challenging problems such as occlusion, illumination change, shadow, and camera motions remained difficult problems that are actively studied today. One relatively under-studied problem is the quality of videos. Current research on video have focused on high-definition videos that offer tremendous details and strong fidelity of signal. However, most of these videos are not feasible for real-time video processing, streaming data and mobile applications, particularly when additional processing is required for the recognition of actions in video.

Visual recognition approaches for images has recently been extended for use in video sequences, with good measures of succes. Particularly, bag-of-features (or bag-of-visual-words) based methods have also shown excellent results for action recognition [1]–[3]. Despite recent developments, the representation of local regions in videos is still an open field of research. For representation of videos, different spatio-temporal features have been considered in literature. Many popular works [1], [4], [5] prefer utilizing gradient and flow information to describe the shape and motion that lies in the video. The use of textures are less common [6], [7], though there are promising benefits that can be leveraged.

Oh et al. [8], in establishing the recent large-scale VIRAT dataset for continuous surveillance, provided nine different downsampled versions of the data in the initial version[1]), consisting of three spatial scales and three temporal frame rates. The authors note that this is a "relatively unexplored area" and that "it is important to understand how existing approaches will behave differently".

Motivated by the known merits of different features and the lack of work in low quality videos, we aim to investigate and present viable approaches to this problem. In this paper, we propose a joint utilization of shape, motion and texture features for robust recognition of human actions in low quality downsampled videos. This idea of representation integrates these well-established feature methods in a new way that alleviates their individual shortcomings. We also investigate and analyze the performance of action recognition reacts under two low quality conditions – spatial downsampling and temporal downsampling. We conduct an extensive set of experiments on two benchmark action datasets, the KTH and Weizmann, both of which are already low in frame resolution in its original form. Finally, the viability of our proposed approach is further analyzed, providing insights into good combination of features and the importance of using kernels to provide a balanced set of features that fit well to the data.

A. Related Work

Human action recognition has been studied extensively in recent years [9]. From the the recent research in activity recognition roughly, spatio-temporal video features can be categorized into three main different categories based on the nature of feature used for classification: dynamic feature (motion), structure (shape) and texture, or implicit or explicit combination of three. Most recent works employ primarily motion and shape features [3]. Laptev [10] first proposed the extraction of shape (HOG) and motion (HOF) information from spatio-temporal interest points (STIP) to classify human actions in video. More recently, Wang et al. [5] proposed the use of dense trajectories with the same way of encoding the shape and motion information. All these methods appear to suggest that the combination of shape and motion features performs better than using them alone.

Spatio-temporal texture features such as LBP-TOP [11] have also found their way to action recognition. Kel-

[1]As of today, these downsampled versions are no longer available in the current VIRAT version 2.0. Website: http://www.viratdata.org/

lokumpu et al. [6] proposed the use of LBP-TOP descriptor to recognize human actions by applying it on the entire bounding volume area. Mattivi and Shao [7] applied LBP-TOP over small video patches called cuboids which are extracted from each interest point, resulting in a more sparse representation of video sequences. Their approach managed a promising accuracy rate of around 91% on the KTH dataset.

II. Spatio-Temporal Video Features

In the following sections we describe the three types of spatio-temporal features that can be extracted from action videos, namely structural (shape), dynamic (motion) and textural (texture) features. As structural and dynamic features are somewhat related, we shall describe them together in Section A, while textural feature is elaborated in Section B.

A. Structural and Dynamic Features

Generally speaking, structural information in video embodies the geometrical or shape-oriented variations found spatially; dynamic information in video carries important temporal information or changes of its structure across time. These two forms of information are typically taken together to exemplify spatio-temporal information in video.

For each given sample point (x, y, t, σ, τ), a feature descriptor is computed for a 3-D video patch centered at (x, y, t) at spatial and temporal scales σ, τ. In this work, we employ the Harris3D detector (a space-time extension of the popular Harris detector [12]) to obtain spatio-temporal interest points (STIP) [10]. Briefly, a spatio-temporal second-moment matrix is computed at each video point $\mu(.;\sigma;\tau) = g(.;s\sigma;s\tau) * (\nabla L(.;\sigma;\tau)L(.;\sigma;\tau))^T$ using a separable Gaussian smoothing function g, and space time gradients ∇L. The final location of the detected STIPs are given by local maxima of $H = \det(\mu) - k\text{trace}^3(\mu)$ [3]. We used the original implementation available online and standard parameter settings i.e. $k = 0.00005$, $\sigma^2 = \{4, 8, 16, 32, 64, 128\}$ and $\tau^2 = \{2, 4\}$, for original videos and a majority of downsampled videos.Figure 1 shows the Harris3D detector being used to extract STIPs on the KTH dataset.

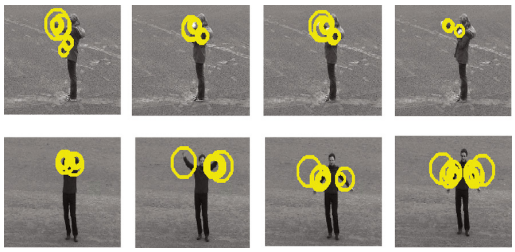

Fig. 1. Harris3D feature detector on KTH data set

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) descriptors as proposed by Laptev in [10]. The combination of HOG/HOF descriptors with interest point detectors produces descriptors of size $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma, \Delta_t(\tau) = 8\tau$. 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 [3]. In this experiment we opted for grid parameters $n_x, n_y = 3, n_t = 2$ for all videos, as suggested in the original paper.

B. Textural Features

Textures are defined as statistical regularities over both space and time, e.g. motion of birds in a flock which was recently used for action recognition with good results. [7].

One of the most widely-used texture descriptor, Local Binary Pattern (LBP) produces a binary code at each pixel location by thresholding pixels within a circular neighborhood region by its center pixel [13]. 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. After computing these LBP patterns for the whole image, an occurrence histogram is constructed to provide a statistical description of the distribution of local textural patterns in the image. This descriptor has been proved to be successful in face recognition [14].

In order to be applicable in the context of dynamic textures such as facial expressions, Zhao et al. [11] proposed LBP on Three Orthogonal Planes (LBP-TOP), where LBP is performed on the three orthogonal planes (XY, XT, YT) in the video volume by concatenating their respective occurrence histograms into a single histogram. LBP-TOP is formally expressed by $LBP - TOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_Z}$ where the subscripts denote a neighborhood of P points equally sampled on a circle of radius R on XY, XT and YT planes respectively. The resulting feature vector is $3 \cdot 2^P$ in length. Fig. 2 illustrates the construction of the LBP-TOP descriptor. As can be seen, LBP-TOP encodes the appearance and motion along three directions, incorporating spatial information in XY-LBP and spatial temporal co-occurrence statistics in XT-LBP and YT-LBP. In this experiment we apply the parameter settings of $LBP - TOP_{8,8,8,2,2,2}$ with non-uniform patterns as specified by Mattivi and Shao [7], which produces a feature vector length of 768.

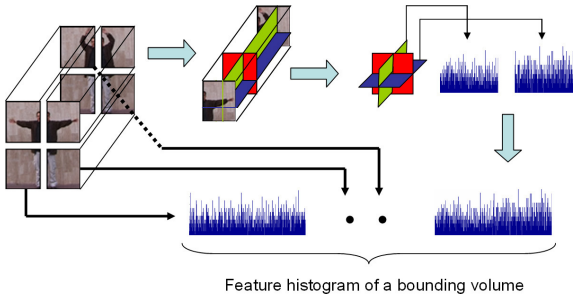

Fig. 2. LBP-TOP feature descriptor. Image from [6]

III. Video Downsampling

A video's spatial resolution and temporal sampling rate defines the amount of spatial and temporal information it can convey. Spatial resolution is simply the video's horizontal pixel count by its vertical pixel count, i.e. frame size. The temporal sampling rate defines the number of discrete frames in a unit of time, i.e. frames per second (fps) or Hertz (Hz).

In this work, we investigate the performance of action recognition with low quality videos that have been downsampled spatially or temporally, proposing suitable features

that are robust. For now, we first describe the spatial and temporal downsampling modes that were employed in this work.

A. Spatial Downsampling

Spatial downsampling produces an output video with a smaller resolution than the original video. In the process, no additional data compression is applied while the frame rates remained the same. For clarity, we define a spatial downsampling factor, α which indicates the factor in which the original spatial resolution is reduced. In this work, we fixed $\alpha = \{2, 3, 4\}$ for modes SD_α, denoting that the original videos are to be downsampled to half, a third and a fourth of its original resolution respectively. Fig. 3 shows a sample video frame that undergoes SD_2, SD_3 and SD_4. We opted not to go beyond $\alpha = 4$ as extracted features are too few and sparse to provide any meaningful representation.

Fig. 3. Spatially downsampled videos. (a) Original (SD_1); (b) SD_2; (c) SD_3; (d) SD_4;

B. Temporal Downsampling

Temporal downsampling produces an output video with smaller temporal sampling rate (or frame rate) than the original video. In the process, the video frame resolution remained the same. Likewise, we also define a temporal downsampling factor, β which indicates the factor in which the original frame rate is reduced.

It has been seen that high temporal resolution; with high spatial resolution produces high dynamic range i.e. high motion information. It is based on the assumption that non-constant intervals would yield jerky motion, i.e. perceivable discontinuity in the optical flow field. This assumption is true for the majority of video sequences, which contain motion, captured at the frame rate of 30 or less. Low quality videos usually have this kind of motion discontinuity.

In this work, we use values of $\beta = \{2, 3, 4\}$ for modes TD_β, denoting that the original videos are to be downsampled to half, a third and a fourth of its original frame rate respectively. In the case of videos with slow frame rates or short video lengths (such as in the Weizmann dataset [15]), β may only take on smaller range of values to extract sufficient features for representation.

IV. Experiments

In this section, we describe a set of extensive experiments and their respective results, while analyzing and comparing different combination of feature descriptors discussed earlier. Experiments were conducted separately for spatial downsampling and temporal downsampling to demonstrate the strengths of specific features with respect to each condition. We also provide a detailed elaboration of the evaluation framework and settings used for the different experimented datasets.

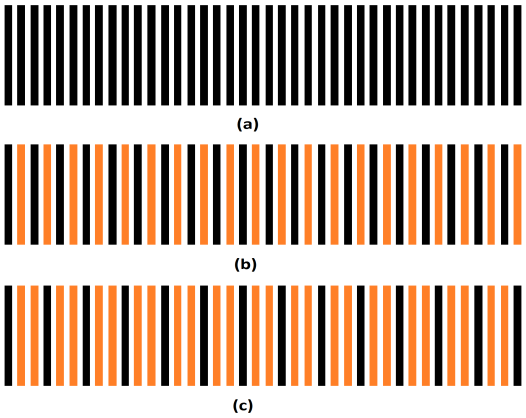

Fig. 4. Temporal Downsampling; (a) Original video (b) TD_2; (c) TD_3;

A. Datasets

We have conducted our experiments on two notable action recognition datasets – the KTH actions dataset [16] and the Weizmann dataset [15]. Both datasets are similar in the way that they are captured in a controlled environment with homogeneously uniform background.

KTH is the most popular dataset in literature for human action recognition. It contains 6 action classes: walking, running, jogging, hand-waving, hand-clapping and boxing; performed by 25 actors in 4 different scenarios: outdoors, outdoors with scale variation, outdoors with different cloths and indoors. There are 599 video samples in total (one subject has less one clip). Each clip is sampled at 25 fps and lasts between 10–15 seconds with image frame resolution of 160×120 pixels. We follow the original experimental setup, i.e., divide the samples into test set (9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22) and training set (the remaining 16 subjects) [16], while reporting the average accuracy over all classes as performance measure.

The **Weizmann** dataset was introduced by Blank et al. [15]. It contains 93 video clips from 9 different subjects (3 subjects have one extra clip) with each video clip containing one subject performing a single action. There are 10 different action categories: walking, running, jumping, gallop sideways, bending, one-hand waving, two-hands waving, jumping in place, jumping jack, and skipping. Each clip lasts about 2–3 seconds at 25 fps (interlaced) with image frame resolution of 180×144 pixels. Testing is performed by *leave-one-person-out cross-validation* (as suggested in [4]) i.e., for each fold, training is done on 8 subjects and testing on all videos of the remaining held-out subject.

B. Evaluation Framework

A video sequence is represented as a bag of local spatio-temporal features [16]. Spatio-temporal features are first quantized into visual words and a video is then represented as the frequency histogram over the visual words. In our experiments, vocabularies are constructed with standard k-means clustering with the number of visual words empirically set to $K = 2000$ to obtain a reasonably good performance across datasets. To limit the complexity, we cluster a subset of 100,000 randomly selected training features. To increase precision, we initialize k-means 8 times and kept the result with the lowest error. Features are assigned to their closest vocabulary word using Euclidean distance. The

resulting histograms of visual word occurrences are used as video sequence representations.

For classification, we use a non-linear support vector machine (SVM) [17] with a χ^2-kernel.

$$K(H_i, Hj) = \exp(-\frac{1}{2A}\sum_{n=1}^{K}\frac{(h_{in}-h_{jn})^2}{h_{in}-h_{jn}})$$

which was previously found to be effective for action recognition [1]. Here, h_{in} and h_{jn} are the frequency histograms of the n-th word occurrences, K is the vocabulary size, and A is the mean value of distances between all training samples [18]. In some parts of our experiments, we also tested with a linear kernel instead of χ^2 kernel, which is known to over-fit the feature data occasionally at higher dimensionality. For multi-class classification, we apply the *one-against-rest* approach and select the class with the highest score.

C. Experimental Results

In this subsection we present the experimental results in three parts, based on the original videos, spatially downsampled videos and temporally downsampled videos. For each part, we systematically compare and analyze the performance of different feature descriptors, providing further insights into the intuition behind the different feature types. Experiments were conducted on an Intel Core-i7 3.6 GHz machine with 24GB RAM.

For ease of reporting, we will compare the following combination of features and settings in all experiments, denoted as follows: **I**: STIP; **II**: STIP-χ^2; **III**: STIP + LBP-TOP; **IV**: STIP + LBP-TOP-χ^2; **V**: (STIP + LBP-TOP)-χ^2. The HOG, HOF and HOG + HOF descriptors will be used on the extracted STIPs, while LBP-TOP is applied on the entire video volume. For features **III**, **IV** and **V**, the STIP-based descriptors are concatenated with the LBP-TOP at the histogram level.

1) Experiments on Original Videos: On the KTH dataset, we obtained the best result of 94.91% using the combination of HOG and HOF features (HOG+HOF) (see Figure 5), which constitutes a histogram-level concatenation of HOG and HOF as opposed to a descriptor-level concatenation (HOGHOF) advocated in [1], [3]. Figure IV-C1 shows that this clearly helps to elevate the overall accuracy by 3–8%. However, on the Weizmann dataset (see Figure 6) , we observe that there is less distinction between the three tested features, with the HOF holding a slight advantage in terms of performance. The best result of 94.44% was achieved using HOF feature.

For both datasets, we also observed that kernelization of specific features are able to strengthen results. For instance on the KTH, HOF + LBP-TOP with an already impressive 93.06% accuracy, is even higher at 94.44% after kernelizing the LBP-TOP features. This is most apparent when LBP-TOP features are kernelized (see Figure IV-C1). Other features in consideration also show similar characteristic except for HOF, which has negligible difference.

In short, dynamic feature (HOF) is notably essential for effective action recognition on the original video samples. Shape feature (HOG) is largely poor on all combinations, but improves tremendously when paired with textural feature (LBP-TOP).

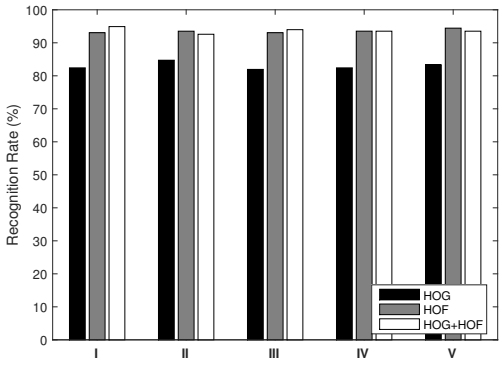

Fig. 5. Recognition rate of different combination of features on original KTH dataset videos

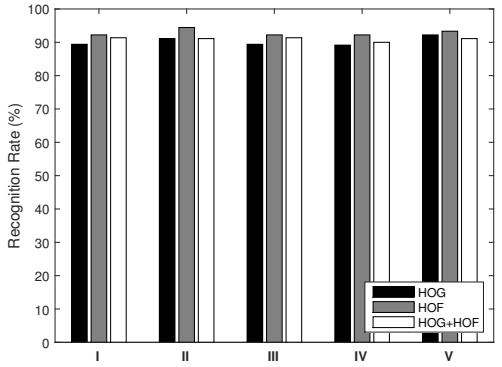

Fig. 6. Recognition rate of different combination of features on original Weizmann dataset videos

2) Experiments on Spatially Downsampled Videos: Table I shows the recognition rate of the five descriptor combinations (**I**-**V**) with different STIP descriptors, on the KTH dataset. Overall, the combination of STIP descriptors + kernelized LBP-TOP appear to dominate the best results within each mode. This clearly shows the important role of motion and textural information with respect to deterioration of spatial quality. As expected, shape information becomes less discriminant as spatial resolution decreases. More promisingly, LBP-TOP contributes significantly more (comparing **IV** to **I** and **II**) as the resolution quality decreases. Nevertheless, it performed well on the Weizmann dataset but not on the KTH datasetwhen used entirely alone (see Figures 9 and 10).

Combinations **IV** and **V** are the two most robust methods, where the STIP descriptors (particularly the HOF feature) are combined with LBP-TOP to great effect; the kernelized LBP-TOP achieving 87.5% accuracy rate at $\alpha = 4$. STIPs were extracted with $k = 0.0001, 0.000075$ and 0.00005 for SD_2, SD_3 and SD_4 respectively to ensure maximum number of interest points with respect to spatial size.

3) Experiments on Temporally Downsampled Videos: Both the KTH and Weizmann datasets have a frame rate of 25 fps; upon downsampling, TD_1: 12.5 fps, TD_2: 8.33 fps and TD_3: 6.25 fps. Table 1 summarizes the recognition rate of the five descriptor combinations (**I** and **V**) with different STIP descriptors, on the KTH dataset. Similarly,

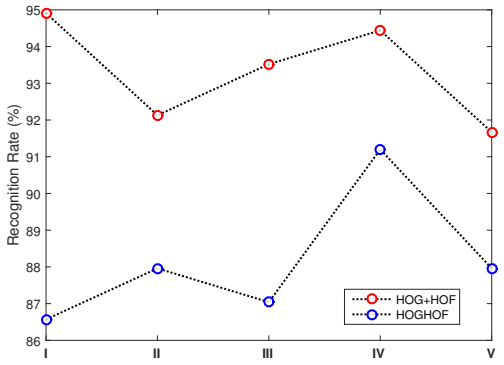

Fig. 7. Comparison between recognition performance of HOG+HOF (histogram-level concatenation) and HOGHOF (descriptor-level concatenation)

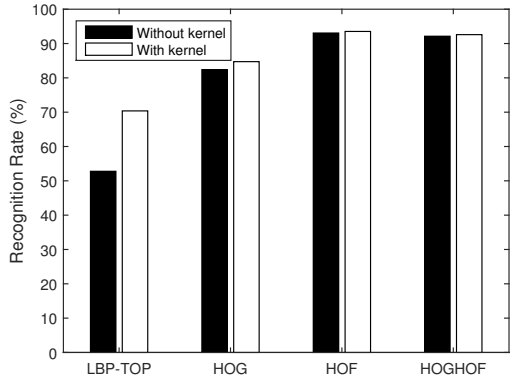

Fig. 8. Recognition accuracy with and without χ^2-kernel, on the original KTH videos.

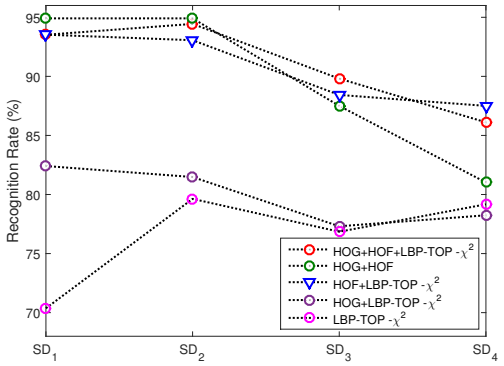

Fig. 9. Performance of selected combination of different features across spatial downsampling modes for KTH dataset

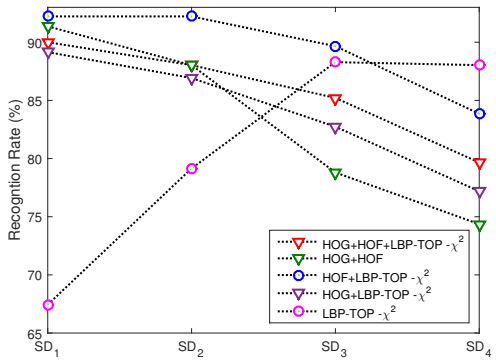

Fig. 10. Performance of selected combination of different features across spatial downsampling modes for Weizmann dataset

TABLE I. RECOGNITION RATE (%) OF VARIOUS DESCRIPTOR COMBINATIONS FOR SPATIALLY DOWNSAMPLED KTH VIDEOS

Mode	Combination	HOG	HOF	HOG+HOF
SD_2	**I**	68.06	91.67	**94.91**
	II	77.22	92.13	92.13
	III	67.59	92.59	93.52
	IV	**81.48**	**93.06**	94.44
	V	75.46	90.74	91.67
SD_3	**I**	62.50	87.04	87.50
	II	62.50	85.65	85.19
	III	62.50	87.04	87.50
	IV	**77.31**	**88.43**	**89.81**
	V	71.76	86.57	85.19
SD_4	**I**	56.94	81.94	81.20
	II	57.94	80.56	82.87
	III	62.50	82.87	81.02
	IV	**78.24**	**87.50**	**86.11**
	V	69.44	83.8	84.26

TABLE II. RECOGNITION RATE (%) OF VARIOUS DESCRIPTOR COMBINATIONS FOR TEMPORALLY DOWNSAMPLED KTH VIDEOS

Mode	Combination	HOG	HOF	HOG+HOF
TD_2	**I**	76.39	87.04	91.20
	II	**80.56**	86.11	89.81
	III	75.00	88.89	91.20
	IV	80.09	**89.81**	**92.59**
	V	79.17	87.04	91.20
TD_3	**I**	68.06	76.85	82.41
	II	**75.46**	77.31	84.26
	III	74.07	78.24	**86.11**
	IV	**75.46**	**82.87**	85.19
	V	73.15	79.63	82.87
TD_4	**I**	66.67	71.76	**82.41**
	II	73.15	73.15	81.94
	III	69.44	73.61	77.78
	IV	**74.04**	**75.46**	**82.41**
	V	72.69	69.44	81.48

we see a strong showing when LBP-TOP is incorporated with around 3–6% improvement in accuracy. Again, this demonstrates the importance of textural information when temporal sampling rate is poor. On the KTH, we see that the use of all three features (shape, motion and texture) promotes robustness against deterioration of temporal quality (Figure 11). Method **IV** commands a respectable 82.41% accuracy rate at $\beta = 4$.

It is also worth mentioning that shape information becomes increasingly useful with the reduction of frame rate (particularly for TD_4 for KTH) since dynamic information becomes more sparse and disjointed. Figures 11 and 12 show the performance of selected feature combinations for different downsampling modes on the KTH and Weizmann dataset respectively.

4) Future Directions: Based on this preliminary work and the analysis of the results obtained, there are several possible directions for future work.

We intend to extend our evaluation to videos from more complex and uncontrolled environments [1], [8]. While

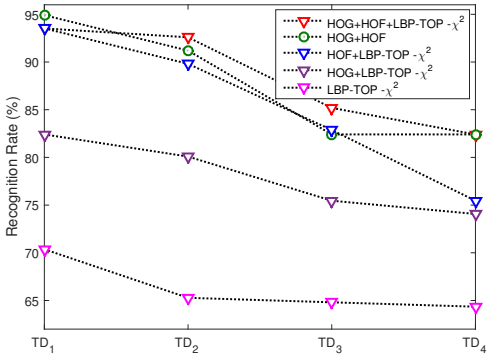

Fig. 11. Performance of selected combination of different features across temporal downsampling modes for KTH dataset

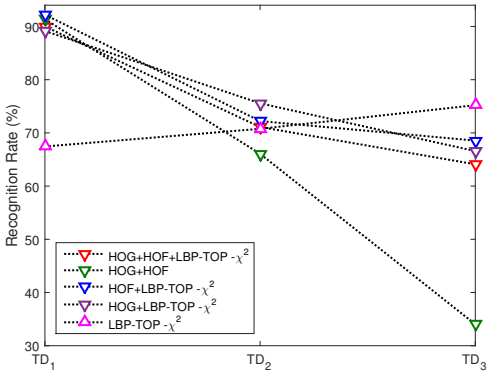

Fig. 12. Performance of selected combination of different features across temporal downsampling modes for Weizmann dataset

our experiments already point towards the sensitivity of different features (shape information is sensitive towards resolution, motion information is sensitive towards sampling/frame rate), it will be interesting to investigate the simultaneous effects of both spatial and temporal downsampling. How well can textural features prop up the recognition capability? Also, the use of LBP-TOP in this work merely illustrates the potential benefits of spatio-temporal texture descriptors in general. We intend to explore other spatio-temporal textural features that might exhibit more robustness towards video quality.

V. Conclusion

In this paper, we explore a new notion of jointly using shape, motion and texture features for action recognition in low quality videos. To the best of our knowledge, there are no existing systematic attempts to investigate the problem of video quality, which is most relevant in many consumer applications and real-life scenarios. This preliminary work draws interesting conclusions on how spatially and temporally downsampled videos can particularly benefit from textural information, considering that most common approaches involved only structural and dynamic information. The combined usage of all three features (HOG+HOF+LBP-TOP) outperforms the other competing methods across a majority of cases. Our best method is able to limit the drop in accuracy to around 8-10% when the video resolutions and frame rates deteriorate to a fourth of their original values.

VI. Acknowledgment

This work is supported, in part, by the Ministry of Education, Malaysia under Fundamental Research Grant Scheme (FRGS) project FRGS/2/2013/ICT07/MMU/03/4.

References

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE CVPR*, 2008, pp. 1–8.

[2] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.

[3] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, pp. 124–1.

[4] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. of the 15th Int. Conf. on Multimedia*. ACM, 2007, pp. 357–360.

[5] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE CVPR*, 2011, pp. 3169–3176.

[6] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Human activity recognition using a dynamic texture based method." in *BMVC*, 2008.

[7] R. Mattivi and L. 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.

[8] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *IEEE CVPR*, 2011, pp. 3153–3160.

[9] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

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

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

[12] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of 4th Alvey Vision Conference*, vol. 15, 1988, p. 50.

[13] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[14] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. PAMI*, vol. 28, no. 12, pp. 2037–2041, 2006.

[15] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *IEEE ICCV*, vol. 2, 2005, pp. 1395–1402.

[16] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Int. Conf. on Pattern Recognition*, vol. 3, 2004, pp. 32–36.

[17] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. of the Int. Conf. on Multimedia*. ACM, 2010, pp. 1469–1472.

[18] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.