Histogram of Fuzzy Local Spatio-Temporal Descriptors for Video Action Recognition

Zheming Zuo, *Student Member, IEEE,* Longzhi Yang, *Senior Member, IEEE,* Yonghuai Liu, *Senior Member, IEEE,* Fei Chao, *Member, IEEE,* Ran Song, *Member, IEEE,* Yanpeng Qu, *Member, IEEE*

Abstract—**Feature extraction plays a vital role in visual action recognition. Many existing gradient-based feature extractors, including histogram of oriented gradients (HOG), histogram of optical flow (HOF), motion boundary histograms (MBH), and histogram of motion gradients (HMG), build histograms for representing different actions over the spatio-temporal domain in a video. However, these methods require to set the number of bins for information aggregation in advance. Varying numbers of bins usually lead to inherent uncertainty within the process of pixel voting with regard to the bins in the histogram. This paper proposes a novel method to handle such uncertainty by fuzzifying these feature extractors. The proposed approach has two advantages: i) it better represents the ambiguous boundaries between the bins and thus the fuzziness of the spatio-temporal visual information entailed in videos, and ii) the contribution of each pixel is flexibly controlled by a fuzziness parameter for various scenarios. The proposed family of fuzzy descriptors and a combination of them were evaluated on two publicly available datasets, demonstrating that the proposed approach outperforms the original counterparts and other state-of-the-art methods.**

Index Terms—**Video feature extraction, histogram, local feature descriptors, fuzziness, action recognition.**

I. INTRODUCTION

VISUAL action recognition systems identify human motion activities in diverse scenes. Intelligent video systems [1] have been intensively studied in recent years due to the wide range of real-world applications such as automated video surveillance for health-care and human-robot interaction. Nevertheless, the classification of realistic unconstrained videos remains as one of the most challenging problems in computer vision, owing greatly to the high degree of spatio-temporal visual variations, the low-resolution video clips, and the various backgrounds and environments. Feature extraction discovers the signatures of actions to enable the utilisation of classifiers, which can be derived either locally or globally. Compared with global descriptors, local descriptors are usually robust to illumination, occlusion and other chaotic backgrounds, and sensitive to small variations between actions.

Multiple gradient-based local feature extraction approaches have been developed using histogram representation. HOG

Z. Zuo, and L. Yang are with the Department of Computer and Information Sciences, Northumbria University, UK. Email: longzhi.yang@northumbria.ac.uk. Y. Liu is with the Department of Computer Science, Edge Hill University, UK. F. Chao is with the Department of Computer Science, Xiamen University, UK. R. Song is with the University of Brighton. Y. Qu is with the Information Science and Technology College, Dalian Maritime University, China.

aggregates the gradient responses in a 2-D space [2]; HOF aggregates optical flow responses [3]; MBH uses the gradients over optical flow fields to construct descriptors in horizontal (i.e. MBHx) and vertical (i.e. MBHy) directions [4]. The combination of these feature descriptors leads to the improved dense trajectories (iDT) [5]. Recently, HMG was proposed as a simple yet efficient temporal derivative method [6]. These gradient-based approaches represent videos using the histograms of spatial, temporal, and optical flow gradients, with each bar denoting the accumulated gradient responses in a certain range of video frame directions. These directions are generally termed as bins and using different numbers of bins usually results in different performances, which entails certain fuzziness in the definitions of bins.

This paper proposes a novel fuzzy voting mechanism for the distribution of spatio-temporal gradient responses to bins, to address the uncertainty. This is achieved by systematically representing the aforementioned gradient-based local feature descriptors via a unified framework and *fuzzify* them to allow each pixel to belong to all the bins with corresponding differed partial memberships. Consequently, the uncertainty caused by the number of bins is effectively mitigated by the partial membership which further benefits the action representation.

The major contributions of this work include: 1) proposing the histogram of fuzzy local spatio-temporal descriptors (HFLSTD) to fuzzify the contributions of pixels to bins, 2) fusing the proposed HFLSTD descriptors to construct the improved fuzzy dense trajectories (iFDT), and 3) applying the proposed HFLSTD and iFDT to the datasets UCF-50 [7] and UCF-101 [8], with the support of the popular iDT [5] and 3D convolutional neural network (C3D) [9] for a comparative study in reference to the state-of-the-art techniques.

The rest of this paper is structured as follows. Section II revisits the related works. Section III details the proposed HFLSTD and iFDT. Section IV evaluates the proposed feature descriptors with the support of deep features. Section V concludes the paper and points out future directions of research.

II. BACKGROUND

The bag of visual words (BoVW) or bag of features (BoF) is one of the most acknowledged frameworks for carrying out action recognition tasks, as outlined in Fig. 1, where images or videos are represented as feature vectors. A large number of homologous but different local feature extraction techniques have been proposed with various levels of success in action recognition [2–5, 10]. These techniques utilised the
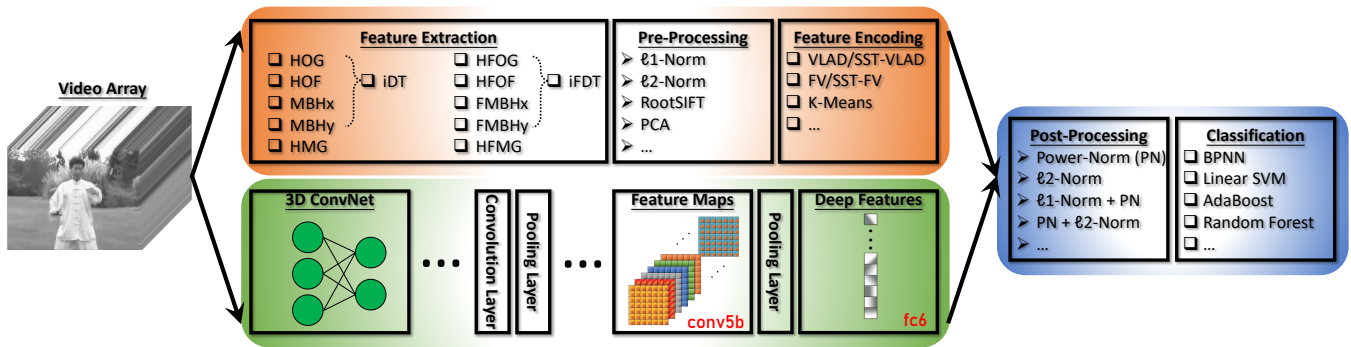
Fig. 1: Video action recognition with local feature descriptors (upper pathway) and deep-feature descriptors (lower pathway).

popular modelling paradigm of histogram as modern feature descriptors. In brief, the local video array is divided into blocks aggregating feature responses within them, followed by their concatenation over the blocks.

HOG descriptor was originally proposed for detecting human bodies from images by concatenating the aggregated two dimensional oriented gradient responses [2], which was later fuzzified for 2-D face recognition [10]. Both the original HOG and its fuzzy extension only capture the spatial or static appearance information, with the temporal information discarded. HOF was proposed for recognising human actions from movies by aggregating 2D optical flow responses [3], whilst the MBH method computes the changes of the optical flow in order to remove the influence of camera motion by calculating horizontal and vertical derivatives on the calculated optical flow from HOF separately (i.e., MBHx and MBHy) [4]. HMG aggregates the magnitude responses in both spatial and temporal derivatives [6], which provides an efficient enhancement of HOG by considering the extra information over time, and an effective alternative of HOF by simplifying the calculation of optical flow. These features are often jointly utilised for improving the performance of visual action recognition [3, 5, 6, 11].

Feature extraction through deep learning has recently been very popular due to the powerful generalisation capability in constructing effective deep features [9, 12–19]. For instance, a 2-D single-frame convolutionary neural network (CNN, ConvNet) model for learning video-wise representation was reported in [12], and a 3D convolutional neural network (C3D) was proposed for recognising human actions spatially and temporally [9], in which the 3D convolution operation was used for obtaining extra temporal dynamics comparing to its conventional 2D counterpart. Also, efficient two-stream spatio-temporal features are constructed by fusing spatial ConvNet and temporal (optical flow) ConvNet interactively [13]. Noting that the C3D and the two-stream CNN are both computationally expensive, a sparse temporal sampling method was proposed in [14] to address this problem, particularly for long videos by randomly picking up a small temporal block in each of the evenly divided video segments. In addition, a second-order temporal pooling strategy was presented in [15] to generate feature descriptors based on temporal classification scores instead of CNN features.

Extracted features can be fused for improving video representation through early or late fusion [5, 6, 9, 11, 13, 19]. Early fusion, or feature-level fusion, concatenates all the local features, or sometimes with deep features, prior to the classification process; late fusion is usually a feature combination weighted by the prediction accuracy. Late fusion can be performed between local features [11], deep features [19], or both [13]. Examples of early fusion approaches include iDT [5] and HMG fused with iDT [6], while examples of late fusion approaches include densely extracted HOG, HOF, MBHx, and MBHy [11]. Recently, the fusion of local features and deep features has shown competitive recognition performance with iDT being a popular candidate. In particular, iDT can be early fused with $\ell 2$ normalised feature vector that is often generated by a fully-connected layer in C3D [9], or late fused with the predictions of two- [13] or four-stream [19] CNN features.

Fuzzy logic has been widely applied in action recognition to represent and reason about source video data with uncertainty [20–22]. To fuse multi-modal features, fuzzy integral was adopted in [20] which nonlinearly fused video and audio features via late fusion. An extended version of extreme learning machine was proposed for multi-view action recognition by tolerating small training errors using the within-class variance of the training instances, where the decorrelation was achieved using fuzzy vector quantisation [21]. A complex-valued interval type-2 neuro-fuzzy inference system was presented in [22] using a Takagi-Sugeno-Kang type fuzzy inference mechanism in the neural networks with Gaussian membership function for improving the prediction accuracy of feature extraction of a video clip.

III. HISTOGRAM OF FUZZY LOCAL SPATIO-TEMPORAL DESCRIPTORS

This section introduces the fuzzified version of the local feature extractors. The resulting fuzzy oriented gradients (HFOG), histogram of fuzzy optical flow (HFOF), fuzzy MBHx (FMBHx), fuzzy MBHy (FMBHy) and histogram of fuzzy motion gradients (HFMG) are collectively termed as histogram of fuzzy local spatio-temporal descriptors (HFLSTD). They share most of the workflow illustrated in Fig. 2. Specifically, they first represent each video clip as a set of small units, namely blocks, as described in Section III-A. Then, following the separate pipelines as demonstrated in Fig. 2, the magnitude
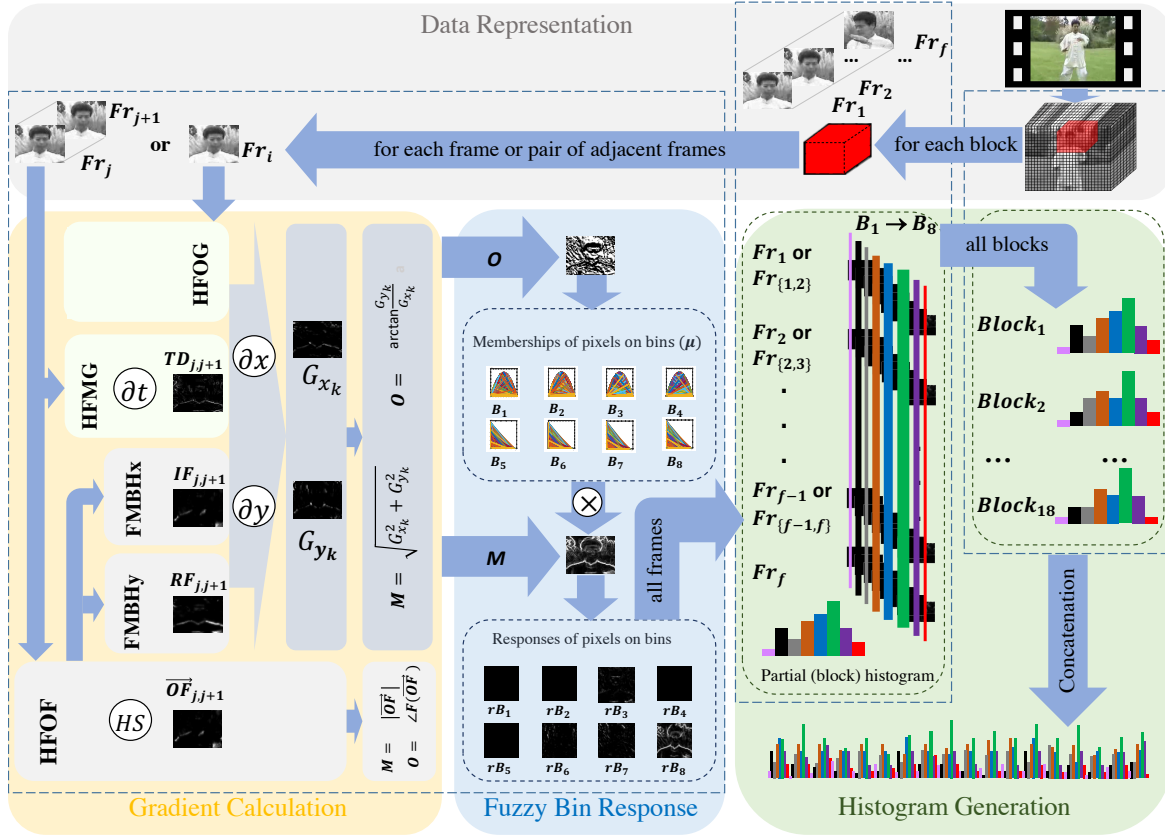
Fig. 2: Outline of the histograms of fuzzy spatial (and temporal) descriptors, based on an example of the pre-defined 18 blocks (3-by-3 spatially by 2 temporally).

and orientation of different types of gradients at each pixel are calculated. Next, a number of bins are designed to accumulate the gradient responses of pixels in different directions (with 8 bins used in Fig. 2 as an example for illustration); and the fuzzy responses of each pixel to all the bins are calculated as detailed in Section III-C. Finally, the responses of all the pixels in each block are aggregated, and then concatenated to form the histogram of gradient-based features to represent the video for action recognition.

A. Information Representation

Each video clip can be viewed as a 3D array of pixels with x and y axes corresponding to an image (i.e. video frame), and t axis representing the time line. Theoretically, gradient-based feature extraction can be performed directly on the pixels in the 3D array. However, practically, by taking the advantages of dense sampling strategy in videos [3] and for better generalisation ability, feature extraction can be accelerated by dividing the 3D array into a number of reusable blocks based on a pre-defined block size. Typically, small block size helps suppress local luminance variations and captures small action movement, while large block size reduces such ability as these variations may be removed by the averaging operation [2]. The selection of block size is usually case-specific, with 8-by-8 pixels spatially and 6 frames temporally most commonly used. Due to the variation of video

sizes, the total number of blocks based on the given block size often varies, and thus some alignment processes, commonly part of the concatenation, were applied [6, 11].

B. Gradient and Orientation Calculation

As the same scheme for feature computation is imposed on every block, a general block consisting of f frames, denoted as $\{Fr_1, Fr_2, \cdots, Fr_f\}$, is taken in this section and Section III-C to facilitate the discussions, as shown in the red box in Fig. 2. Note that the fuzzified feature extractors share exactly the same gradient calculations Gx_j and Gy_j with the corresponding original versions, based on the finite difference between neighbouring frames Fr_j and Fr_{j+1} or pixels p_j along different directions x and y. Different from HFOG and HFMG, HFOF and FMBH both compute the optical flow [23] from every consecutive pair of frames represented as a complex vector $\vec{OF}_{j,j+1}$ with real part $RF_{j,j+1}$ and imaginary part $IF_{j,j+1}$. The calculation of the magnitude M and the direction O for different methods are summarised below:

$$M = \begin{cases} |\vec{OF}| & \text{if HFOF,} \\ \sqrt{Gx_j{}^2 + Gy_j{}^2} & \text{otherwise.} \end{cases} \quad O = \begin{cases} \angle F(\vec{OF}) & \text{if HFOF,} \\ \arctan(\frac{Gy_j}{Gx_j}) & \text{otherwise.} \end{cases}$$
(1)

where $|\vec{OF}|$ represents the complex magnitude, and the phase angle $\angle F(\vec{OF})$ is the orientation of the gradient in radians. As

a consequence, each pixel in the block (except some in block boundaries which are neglectable) is represented by a pair, which expresses the magnitude and direction of its gradient.

C. Fuzzy Bin Response Calculation

The gradient orientations are evenly quantised into b bins in the range of $[0, 2\pi]$, i.e. $B_i = 2\pi \cdot i/b$, $i \in \{0, 1, \cdots, b - 1\}$. The original family of approaches partially assign each pixel to the 2 closest bins, according to how far away it is from them. This might not be sufficient to represent the visual information entailed in the real-world video clips due to the uncertainty about the relationship between the pixel and the two neighbouring bins. This leads to the natural appeal of the development of a fuzzy extension of the gradient measure as discussed. In particular, the fuzzy extension of the family of approaches developed herein assigns a partial membership of each pixel to every bin to address the uncertainty in bin boundaries and affiliations.

Suppose that b bins are used. Given the jth pixel p_j in a block frame with gradient orientation O_j. Let $\mu_{B_i}(p_j)$ represent the pixel intensity of p_j regarding bin B_i, which can be calculated as:

$$\mu_{B_i}(p_j) = \frac{1}{\sum_{k=1}^{b} \left(\frac{\|O_j - O(B_i)\|}{\|O_j - O(B_k)\|} \right)^{\frac{2}{c-1}}}, \qquad (2)$$

where c is a parameter employed for controlling the fuzziness of frame pixel distributions or fuzziness coefficient, which is usually valued greater than 1. As stated earlier, Eq. (2) is an analogy of the approach used in fuzzy c-means (FCM) clustering. The value of c in FCM and its variants have been extensively studied [24]. However, there is still not a widely accepted or agreed approach for the objective determination of the value of c, and this value is usually selected subjectively in practical applications. An empirical study of this parameter is also included in this research as detailed in Section IV.

D. Histogram Generation

Once the membership $\mu_{B_i}(p_j)$ of each pixel p_j in the block $\mathbb{B}$ with regard to every bin B_i is calculated, the block-wise partial histogram, i.e. the bar representing the block response r with respect to magnitude M_{p_j} of p_j and bin B_i, is constructed as:

$$r_{B_i} = \sum_{p_j \in \mathbb{B}} M_{p_j} \cdot \mu_{B_i}(p_j). \qquad (3)$$

The collection of all the bars representing the bins forms the partial histogram representing the block. Then a normalisation process, such as $\ell 2$, is applied to each of the generated partial histograms. Finally, the partial histograms led by all the blocks are concatenated into an overall histogram, i.e. a single feature vector, for the representation of the video. This vector is then pre-processed, as shown in Fig. 1, using normalisation and dimension reduction to decorrelate features, prior to the feature encoding phase which generates the final feature vector in a fixed uniform length to represent the video clip. After post-processing, classifiers are applied for video action recognition.

E. Complexity Analysis on HFLSTD

Algorithm 1: Histogram of Fuzzy Local Spatio-Temporal Descriptors (HFLSTD)

Input   : V: a video clip $V \in \mathbb{R}^{m \cdot n \cdot f}$
             b: pre-defined number of bins
             $[m' \cdot n' \cdot f']$: pre-defined block size
             c: fuzziness coefficient

Output: $\mathcal{F}$: fuzzy histogram of visual features

1: $\mathcal{B}$-**Step:** Build block representation for V based on the block size $[m' \cdot n' \cdot f']$

2: $\mathcal{I}$-**Step:** Compute spatio-temporal information for each pixel p_j in each block using Eq. (1)

3: $\mathcal{R}$-**Step:** Compute pixel-wise response regarding each bin B_i using Eq. (2)

4: $\mathcal{P}$-**Step:** Calculate block-wise partial histogram using Eq. (3)

5: $\mathcal{H}$-**Step:** Obtain $\mathcal{F}$ by concatenating partial histograms of blocks

---

The proposed method for generating local descriptor is summarised in Algorithm 1. The fuzzified feature descriptors and their original counterparts share some common operations, such as the generation of blocks, the construction of a partial histogram for each block and the final histogram generation using concatenation. These common operations are typically neglectable in the computational complexity analysis as these one-off operations constitute only a small part of the overall computation. Therefore, only the operations on all generated blocks, i.e. $\mathcal{I}$-**Step** and $\mathcal{R}$-**Step** in Algorithm 1, are analysed here. It is noteworthy that, in $\mathcal{I}$-**Step**, the block border pixels are ignored due to the lack of neighbouring pixels for gradient calculation. Suppose that the frame resolution of a given action video clip is m-by-n, and the video clip includes f frames. Also, suppose that the predefined block size is m'-by-n' pixels by f' frames. The central part of the video clip is evenly divided into $N = \lfloor n/n' \rfloor \cdot \lfloor m/m' \rfloor \cdot \lfloor f/f' \rfloor$ blocks for processing.

Note that each feature descriptor in the HFLSTD family works on all the blocks. Therefore, the computational complexity for each descriptor is the computational complexity of processing a single block multiplied by the number N of total blocks. Since the computation of the spatial (magnitude and orientation) and the temporal (optical flow) information involves only the neighbouring pixels inside a block between neighbouring frames, all the algorithms considered here have a linear computational complexity of $\mathcal{O}(N)$.

F. Improved Fuzzy Dense Trajectories

As discussed in Section I, HOG, HOF, MBHx, and MBHy are combined as iDT [5] for performance improvement. This combination effectively removes the dense trajectories in the background of video frames led by camera motion. Fuzzified iDT is thus proposed here, termed as improved fuzzy dense trajectories (iFDT) by combining the corresponding four members of HFLSTD, i.e. HFOG, HFOF, FMBHx, and FMBHy.

The proposed iFDT feature descriptor can be computed as the concatenation of HFOG, HFOF, FMBHx and FMBHy.

IV. EXPERIMENTAL RESULTS

The proposed HFLSTD and iFDT were applied to two publicly available datasets for a thorough evaluation and a comparative study. The experiments were carried out on a Dell workstation with AMD® Ryzen™ Threadripper™ 1950X (16-core) CPU @ 3.40 GHz and 64GB RAM. The work was implemented in the environment of MATLAB® 2018b, TensorFlow 1.4.1 with Python 2.7 (for deep feature extraction) using 2 NVIDIA® GeForce® GTX 1080Ti GPUs. The **UCF-50** [7] action recognition dataset includes 6,618 video clips labeled in 50 categories based on the action subjects [7], whilst the **UCF-101** [8] dataset contains 13,320 videos (with various time durations) from real scenes covering 101 action categories [8]. To facilitate the comparative study, by following the work of [6, 17], 10 randomly selected video clips from each category of the UCF-50 dataset (i.e., 500 videos in total) were used for the experiments, and the performance was evaluated via the leave-one-group-out cross-validation; by following the experimental practice in [5, 9, 14, 15, 18, 19], all the data instances in UCF-101 were utilised in the experiments, and the performance was evaluated by the average recognition accuracy over the three official data splits [8].

A. Experiment Setting

1) **Local Feature Extraction, Pre-processing, and Encoding:** Local features were extracted using block sizes of [8×8×6] with a temporal stride of 6 frames for UCF-50, and [32×32×6] with frame sampling rate of 6 frames for UCF-101, to limit memory usage. The implementation of the experimentation followed some of the suggestions given in [5, 9, 14, 19, 25], and the Horn-Schunk method [23] was used for calculating the optical flow. The implementation of HMG in [6] was also adopted. RootSIFT (the square root of $\ell 1$-normalisation) and principal component analysis (PCA) were adopted for pre-processing. When deploying PCA, the dimensionality of the normalised local features were reduced to 72 from 144 dimensions for both datasets. VLAD [26] and FV [27] were combined with sum-pooling [25] for feature encoding in this work, with 512 and 256 centroids, respectively. 500k features were randomly sampled as suggested in [6] for training the visual words under both encoding schemes, and thus they led to the same dimension of the encoded features.

2) **Deep Feature Extraction:** This work also investigated the feature complementarity and assortment issue by combining the proposed local features with deep C3D features. In particular, the local and deep features were combined by utilising the fully-connected layer `fc6`, as illustrated in Fig. 1, from the C3D model; and the model was constructed with 16 layers. The model was pre-trained on the Sports-1M dataset [12] and 16 frames were randomly selected for all the videos in the employed two datasets, leading to a 4096-dimensional vector representing each video. The choice of this layer guarantees efficient early fusion [9, 19] of deep features and the proposed HFLSTD or iFDT which was constructed by following the experimental procedure reported in [5].

3) **Post-processing and Classification:** The encoded local or extracted deep features were further normalised using the the power normalisation plus the $\ell 2$ normalisation in which the power coefficient was fixed to 0.5 in this work. Then, a back-propagation neural network [28] with one hidden layer of 30 hidden neurons was used for classification.

B. Determination of Fuzziness Parameter

The fuzziness parameter c as the only tunable parameter defined in Eq. (2) is difficult to be theoretically determined as discussed in Section III-C, and thereby is usually empirically studied in order to obtain reliable and optimal results. To this end, the effectiveness of different parameter values was investigated first with the results reported in Tables I and II. In particular, parameter values 1.0125, 1.025, 1.05, 1.1, 2.0, 2.5, 2.75, 3.0, 3.25, 3.5, 4.0, 5.0 and 6.0 were applied to the UCF-50 dataset, whilst 1.0125, 1.1, 2.0, and 3.0 were applied to the UCF-101 dataset. The best performance obtained are highlighted in gray in these tables.

The experiments are grouped into two parts in Tables I and II based on the employed feature encoding approaches and datasets. On the whole, it can be seen that there is a non-linear relationship between the performance of different algorithms and the values of c and different values of c can lead different algorithms to have the same performance. In this case, the one with the smallest c value is highlighted due to its slightly higher efficiency. The results led by FV encoding scheme are overall better than those led by VLAD. In the end, $c = 1.0125$ was selected under the encoding scheme of VLAD and $c = 2.5$ under the scheme of FV, unless otherwise stated, for the experiments in the rest of this paper for better performance.

C. Experimental Results

1) Results on UCF-50 and Discussions: The best performances resulted from the fuzzified feature descriptors were compared with those from their original versions, as listed in Table III, and visually summarised in Figs. 3 (a) and (b) with the corresponding confusion matrices for UCF-50 embedded. Overall, all the fuzzified descriptors outperformed their original counterparts if a proper c was applied, with the most significant enhancements achieved by HFOG by as much as 13.8% and 10.2% under the encoding schemes of VLAD and FV, respectively. This achievement was reached as the HFOG utilises only spatial information and the uncertainty caused by the unattended temporal information was well represented and managed by fuzzy logic. The enhancements led by other fuzzified versions were not as great as HFOG, but still statistically significant under both VLAD and FV feature encoding approaches.

To demonstrate the superiority of the proposed fuzzy local spatio-temporal descriptors over the original versions in video action recognition, HFOG with $c = 2.5$ and HOG under the same FV encoding scheme are selected for comparison. The overall accuracy of HOG is 85.2% and that of HFOG is 95.4%. Note that, in this experiment, 10 video clips were randomly selected from every category of the UCF-50 dataset for all 50 categories of human actions for training. Thus, 10

c	VLAD (%)					FV (%)				
	HFOG	HFMG	HFOF	FMBHx	FMBHy	HFOG	HFMG	HFOF	FMBHx	FMBHy
1.0125	95.4	86.4	84.6	84.6	84.8	95.0	86.4	84.0	85.2	84.2
1.025	95.2	85.8	84.6	84.2	84.6	95.2	87.6	83.8	84.8	84.6
1.05	94.8	85.6	84.2	84.6	84.8	95.0	87.0	84.8	85.0	84.6
1.1	85.0	86.0	84.6	83.6	84.6	87.2	87.0	85.2	84.8	85.2
2.0	82.4	76.8	76.6	75.2	74.8	86.6	78.0	79.0	77.0	76.4
2.5	94.4	85.4	83.8	84.2	83.6	95.4	87.4	84.8	85.0	85.6
2.75	94.4	84.6	83.6	84.4	83.4	95.4	86.8	84.2	85.4	85.2
3.0	94.4	85.0	83.4	83.8	83.6	95.2	86.6	85.0	85.6	85.0
3.25	94.2	84.2	83.6	83.6	83.2	95.2	86.0	84.8	84.8	85.2
3.5	94.4	84.8	82.8	83.6	83.2	94.6	86.0	85.2	85.2	85.4
4.0	94.4	84.4	82.4	84.2	83.2	94.6	86.2	85.0	85.2	85.0
5.0	94.4	84.4	83.0	83.6	83.4	94.4	85.8	85.0	85.0	84.6
6.0	94.0	84.8	83.2	83.4	83.4	94.0	85.4	84.6	85.0	84.8

TABLE I: Performance of feature descriptors with various fuzziness coefficient values (c) over the UCF-50 dataset, and the best performance highlighted in grey.
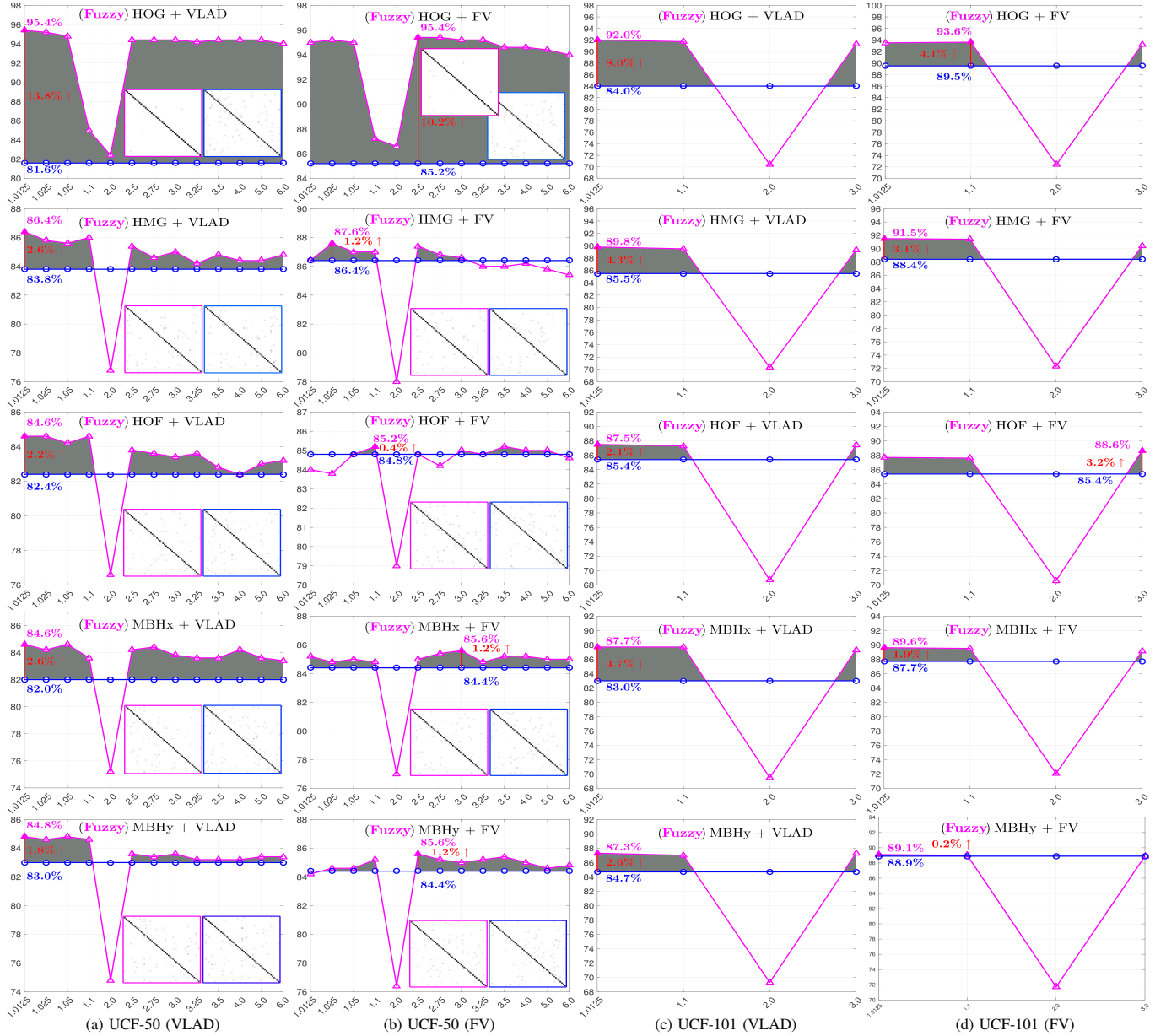

Fig. 3: Accuracy comparison between fuzzy (□ for confusion matrix) and non-fuzzy (□ for confusion matrix) local features on UCF-50 and UCF-101 datasets. The x and y axes represent the fuzziness coefficient (c) and the mean accuracy. The grey region highlights the performance difference between those corresponding fuzzy and conventional feature descriptors, and the optimal c value is marked with ▲.

FeatDesc	VLAD (%)				FV (%)			
	Split1	Split2	Split3	Mean	Split1	Split2	Split3	Mean
HOG	82.3	84.1	85.7	84.0	80.6	93.9	94.0	89.5
HFOG (c@1.0125)	92.0	92.2	91.9	92.0	93.5	93.5	93.4	93.5
HFOG (c@1.1)	91.8	91.6	91.7	91.7	93.6	93.6	93.5	93.6
HFOG (c@2.0)	70.4	70.6	70.2	70.4	72.6	71.6	72.9	72.4
HFOG (c@3.0)	91.4	91.4	91.0	91.3	93.4	93.4	92.7	93.2
HMG	86.8	88.9	80.7	85.5	90.2	84.5	90.5	88.4
HFMG (c@1.0125)	89.9	89.9	89.5	89.8	91.7	91.6	91.2	91.5
HFMG (c@1.1)	89.6	89.8	89.2	89.5	91.5	91.5	91.2	91.4
HFMG (c@2.0)	70.3	70.0	70.5	70.3	71.1	71.3	74.6	72.3
HFMG (c@3.0)	89.5	89.4	89.1	89.3	90.7	90.3	90.3	90.4
HOF	84.5	86.9	84.7	85.4	89.4	83.5	83.3	85.4
HFOF (c@1.0125)	87.9	87.5	87.0	87.5	87.5	87.8	87.9	87.7
HFOF (c@1.1)	87.6	87.4	87.0	87.3	87.2	87.7	87.9	87.6
HFOF (c@2.0)	68.8	68.5	68.7	68.7	70.9	70.4	70.6	70.6
HFOF (c@3.0)	87.6	87.2	87.3	87.4	88.4	88.9	88.5	88.6
MBHx	78.7	83.4	87.0	83.0	84.5	89.8	88.8	87.7
FMBHx (c@1.0125)	88.1	87.9	87.1	87.7	89.7	90.1	89.0	89.6
FMBHx (c@1.1)	88.0	88.0	87.2	87.7	89.6	89.8	89.1	89.5
FMBHx (c@2.0)	69.1	69.6	69.8	69.5	72.9	71.8	71.6	72.1
FMBHx (c@3.0)	87.3	87.4	87.3	87.3	89.0	89.7	88.5	89.1
MBHy	82.3	85.0	86.9	84.7	88.9	88.7	89.1	88.9
FMBHy (c@1.0125)	87.4	87.4	87.2	87.3	89.0	89.2	89.1	89.1
FMBHy (c@1.1)	87.0	87.3	86.6	87.0	89.3	89.1	88.6	89.0
FMBHy (c@2.0)	69.2	69.4	69.4	69.3	72.4	71.0	71.8	71.7
FMBHy (c@3.0)	87.6	87.3	87.0	87.3	89.0	89.1	88.6	88.9

TABLE II: Performance of feature descriptors with various fuzziness coefficient values (c) over the UCF-101 dataset, and the best performance marked in grey.

FeatDesc	VLAD (%)				FV (%)			
	Original	Fuzzified	Var.	Perc.	Original	Fuzzified	Var.	Perc.
HOG	81.6	95.4	13.8	16.9	85.2	95.4	10.2	12.0
HMG	83.8	86.4	2.6	3.1	86.4	87.6	1.2	1.4
HOF	82.4	84.6	2.2	2.7	84.8	85.2	0.4	0.5
MBHx	82.0	84.6	2.6	3.2	84.4	85.6	1.2	1.4
MBHy	83.0	84.8	1.8	2.2	84.4	85.6	1.2	1.4

TABLE III: Classification accuracies of different methods over the UCF-50 dataset, with the same color schemes as used in Fig. 3 with better performance highlighted in **bold**.

predictions were made for each class, with the performance demonstrated in the bar graph in Fig. 4(a). The diagram clearly shows that HFOG outperformed HOG for 35 classes and achieved the same performance with HOG for 13 classes, and underperformed for only 2 classes.
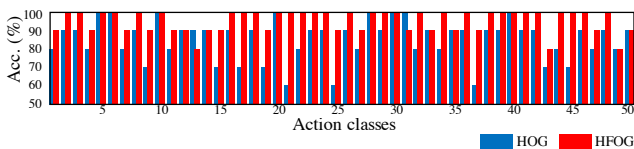
(a) Class perspective
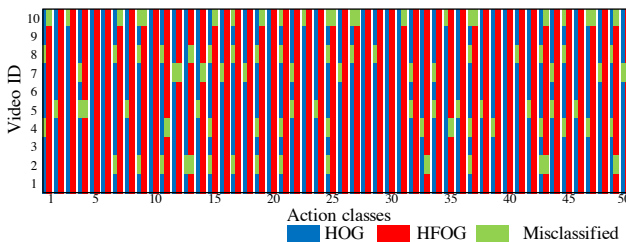
(b) Case perspective

Fig. 4: UCF50 - HOG (85.2%) V.S. HFOG (95.4%) with FV.

More specifically, 413 out of 500 instances were correctly classified by both HFOG and HOG; 10 instances were incorrectly classified by both approaches; and 74 and 23 instances

were misclassified by HOG and HFOG respectively. This implies that 64 cases were successfully classified by HFOG but unsuccessfully by HOG, and 13 instances were correctly classified by HOG but not by HFOG, as summarised in Fig. 4(b). Compared to HOG, HFOG correctly labels some challenging cases as illustrated in the upper two rows of Fig. 5; for instance, 'biking' was misclassified as 'horse riding' and 'clean and jerk' was misclassified as 'pullup'. The cases misclassified by both are often featured with occlusion and chaotic background, with two examples illustrated in the bottom row of Fig. 5.
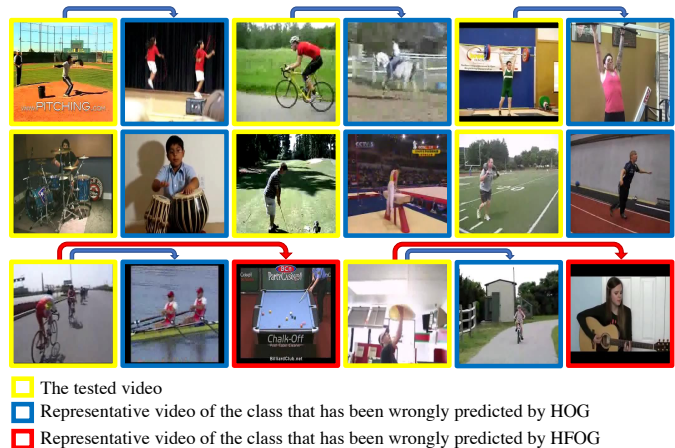

☐ The tested video
☐ Representative video of the class that has been wrongly predicted by HOG
☐ Representative video of the class that has been wrongly predicted by HFOG

Fig. 5: Examples of wrong action predictions using HOG features (top two rows) and both HOG and HFOG features (bottom row) over the UCF-50 dataset.

2) Results on UCF-101 and Discussions: The performances of the competing local feature descriptors are listed in Table IV and visually summarised in Figs. 3(c) and (d). In general, the fuzzified cohort achieved superior performance compared to their original counterparts. Note that all the optimal results generated by the fuzzified ones were based on the same c value of 1.0125 under the VLAD scheme. Similar to the experimental results for UCF-50, the most noticeable improvements were achieved by HFOG for 8% and 4.1% when c took 1.0125 under VLAD and 1.1 under FV. It is followed by HFMG which achieved 89.8% by boosting up 4.3% compared to HMG and 91.5% by boosting up 3.1% using the two encoding methods, with c being 1.025 for both situations. Then the optical flow based three descriptors HFOF, FMBHx, and FMBHy shared similar performance, achieving 87.5%, 87.7%, and 87.3% respectively, whereas FMBHx (89.6% with an improvement of 1.9%) and FMBHy (89.1% that improved by 0.2%) appeared more competitive than HFOF (88.6% with an improvement of 3.2%) when FV was used for feature encoding.

3) Comparison with the state-of-the-art techniques: In this section, the experiments investigate the performance of various combinations of the proposed HFLSTD and iFDT, as well as the state-of-the-art hand-crafted and deep learning feature descriptors, iDT and C3D, as summarised in Table V. The experiments confirmed not only the power of the proposed HFLSTD and iFDT, but also the benefits of combining them

FeatDesc	VLAD (%)				FV (%)			
	Original	Fuzzified	Var.	Perc.	Original	Fuzzified	Var.	Perc.
HOG	84.0	**92.0**	8.0	9.5	89.5	**93.6**	4.1	4.6
HMG	85.5	**89.8**	4.3	5.0	88.4	**91.5**	3.1	3.5
HOF	85.4	**87.5**	2.1	2.5	85.4	**88.6**	3.2	3.7
MBHx	83.0	**87.7**	4.7	5.7	87.7	**89.6**	1.9	2.2
MBHy	84.7	**87.3**	2.6	3.1	88.9	**89.1**	0.2	0.2

TABLE IV: Classification results over the UCF-101 dataset, using the same color code of Fig. 3, with better performance highlighted in **bold**.

with deep features. For example, under the two encoding schemes of VLAD and FV: (i) for the UCF-50 dataset, the HFOG achieved 98.8% and 99.0% when combined with C3D, whilst the iFDT has been boosted up to 96.8% and 96.6%; (ii) for the UCF-101 dataset, fusing iFDT and C3D achieved 97.3% and 97.5% respectively. In particular, classification accuracies of 97.8% and 98.3% are obtained when fusing HFOG with iDT+C3D and C3D under the VLAD and FV respectively. Even though the proposed technique is general and has the advantage of easy implementation, it has achieved even better performance than the state-of-the-art method [17] over the UCF-50 dataset by as much as 2.0%, and Four-Stream Dynamic Image Nets (DIN)+iDT [19] over the UCF-101 dataset by as much as 2.3%. The results have further demonstrated the power of the proposed method and the potential of fuzzy logic in managing uncertainties in the field of computer vision.

UCF-50 (%)		UCF-101 (%)	
CSIFT + MBH [7] (2013)	76.9	Slow fusion [12]† (2014)	65.4
HOG + HOF + MBH(x/y) [11] (2015)	81.8	Hybrid representation [25] (2016)	87.9
MPEG flow + FV [29] (2014)	82.2	C3D (3 nets) + iDT [9]† (2015)	90.4
Causality Descriptor [30] (2014)	89.4	Fixed Model Reuse [16]† (2017)	91.6
iDT + Human Detection + FV [5] (2016)	91.7	$LTC_{Flow+RGB}$ + iDT [18]† (2018)	92.7
Hybrid representation [25] (2016)	92.3	Two-Stream fusion + iDT [13]† (2016)	93.5
Fixed Model Reuse [16]† (2017)	92.4	ST-VLMPF + HMG + iDT [17]† (2017)	94.3
HMG + iDT [6] (2016)	93.0	Two-Stream TSN [14]† (2018)	94.9
MIFS [31] (2015)	95.4	BKCP + KCP + iDT + FV [15]† (2019)	95.4
ST-VLMPF + HMG + iDT [17]† (2017)	97.0	Four-Stream DIN + iDT [19]† (2018)	96.0
HFOG (c@1.0125, VLAD)	**95.4**	**HFOG (c@1.0125, VLAD)**	**92.0**
HFOG + iDT	**95.8**	**HFOG + iDT**	**93.7**
HFOG + C3D†	**98.8**	**HFOG + C3D†**	**97.4**
HFOG + iDT + C3D†	**98.4**	**HFOG + iDT + C3D†**	**97.8**
HFOG (c@2.5, FV)	**95.4**	**HFOG (c@1.1, FV)**	**93.6**
HFOG + iDT	**96.0**	**HFOG + iDT**	**94.7**
HFOG + C3D†	**99.0**	**HFOG + C3D†**	**98.3**
HFOG + iDT + C3D†	**97.6**	**HFOG + iDT + C3D†**	**98.0**
iFDT (c@1.0125, VLAD)	**92.8**	**iFDT (c@1.0125, VLAD)**	**94.0**
iFDT + iDT	**94.0**	**iFDT + iDT**	**95.0**
iFDT + C3D†	**96.8**	**iFDT + C3D†**	**97.3**
iFDT + iDT + C3D†	**96.2**	**iFDT + iDT + C3D†**	**97.2**
iFDT (c@2.5, FV)	**93.6**	**iFDT (c@1.1, FV)**	**94.9**
iFDT + iDT	**94.0**	**iFDT + iDT**	**95.8**
iFDT + C3D†	**96.6**	**iFDT + C3D†**	**97.5**
iFDT + iDT + C3D†	**95.4**	**iFDT + iDT + C3D†**	**97.3**

TABLE V: Classification accuracies of different methods over the UCF-50 and UCF-101 datasets, with deep learning approaches marked with †.

V. CONCLUSION

The paper introduces the idea of *fuzzifying* a group of local spatio-temporal descriptors, including HFOG, HFOF, FMBHx, FMBHy, HFMG, and their combination, iFDT, to handle the uncertainty caused by the construction of histogram-like local features. This is achieved by allowing partial memberships of pixels regarding all bins. The experimental results using the two publicly accessible datasets UCF-50 and UCF-101

demonstrate the efficacy of the fuzzified local descriptors in video action recognition tasks. The proposed descriptors in general outperform the state of the art, despite simple architectures and thus easy implementation. Such results show that the fuzzification of the votes of pixels to all the bins in the histograms is powerful for fusing the information from different frames and representing different actions. The performance of the proposed approaches could be further enhanced by employing different classification methods in conjunction with other pre- and post-processing strategies as well as improved feature encoding schemes, which remain as the future work.

REFERENCES

[1] H. Liu, S. Chen, and N. Kubota, "Intelligent video systems and analytics: A survey," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1222–1233, 2013.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[4] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 428–441.

[5] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *Int. J. Comput. Vision*, vol. 119, no. 3, pp. 219–238, Sep. 2016.

[6] I. C. Duta, J. R. R. Uijlings, T. A. Nguyen, K. Aizawa, A. G. Hauptmann, B. Ionescu, and N. Sebe, "Histograms of motion gradients for real-time video classification," in *Proc. IEEE Int. Work. Content. Mul. Indexing*, 2016, pp. 1–6.

[7] K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vision Appl.*, vol. 24, no. 5, pp. 971–981, 2013.

[8] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," ser. CRCV-TR-12-01. University of Central Florida, 2012.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[10] A. I. Salhi, M. Kardouchi, and N. Belacel, "Histograms of fuzzy oriented gradients for face recognition," in *Int. Conf. Comput. Appl. Technol.*, 2013, pp. 1–5.

[11] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, "Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off," *Int. J. Multimed. Inf. Retr.*, vol. 4, no. 1, pp. 33–44, 2015.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1725–1732.

[13] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[14] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[15] A. Cherian and S. Gould, "Second-order temporal pooling for action recognition," *Int. J. Comput. Vision*, vol. 127, no. 4, pp. 340–362, 2019.

[16] Y. Yang, D. C. Zhan, Y. Fan, Y. Jiang, and Z. H. Zhou, "Deep learning for fixed model reuse," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2831–2837.

[17] I. C. Duta, B. Ionescu, K. Aizawa, and N. Sebe, "Spatio-temporal vector of locally max pooled features for action recognition in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3205–3214.

[18] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, June 2018.

[19] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2799–2813, Dec. 2018.

[20] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, "Realistic human action recognition with multimodal feature selection and fusion," *IEEE Trans. Syst., Man, Cybern. Syst*, vol. 43, no. 4, pp. 875–885, 2013.

[21] A. Iosifidis, A. Tefas, and I. Pitas, "Minimum class variance extreme learning machine for human action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1968–1979, 2013.

[22] K. Subramanian, R. Savitha, and S. Suresh, "A metacognitive complex-valued interval type-2 fuzzy inference system," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1659–1672, 2014.

[23] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1-3, pp. 185–203, 1981.

[24] Y. Zhang, X. Bai, R. Fan, and Z. Wang, "Deviation-sparse fuzzy c-means with neighbor information constraint," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 1, pp. 185–199, Jan. 2019.

[25] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understand.*, vol. 150, pp. 109–125, 2016.

[26] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[27] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[28] J. de Villiers and E. Barnard, "Backpropagation neural nets with one and two hidden layers," *IEEE Trans. Neural Netw.*, vol. 4, no. 1, pp. 136–141, 1993.

[29] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2593–2600.

[30] S. Narayan and K. R. Ramakrishnan, "A cause and effect analysis of motion trajectories for modeling actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2633–2640.

[31] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 204–212.

Yonghuai Liu (SM'11) is a full professor at Edge Hill University since 2018. He is currently an area/associate editor, and editorial board member for various international journals and conference proceedings such as Pattern Recognition Letters, Neurocomputing, American Journal of Educational Research, and IEEE International Conference on Robotics and Automation (ICRA). He won several international awards such as the best associate editor award as an associate editor for the Proceedings of ICRA, 2017. He has published four books and more than 180 papers in top tanked international conference proceedings and journals. His primary research interests lie in 3D computer vision, image processing, pattern recognition, machine learning, artificial intelligence, and intelligent systems. He is a senior member of IEEE, Fellow of British Computer Society and Fellow of Higher Education Academy of United Kingdom.

Fei Chao (M'11) received the B.Sc. degree in Mechanical Engineering from the Fuzhou University, China, and the M.Sc. Degree with distinction in computer science from the University of Wales, Aberystwyth, U.K., in 2004 and 2005, respectively, and the Ph.D. degree in robotics from the Aberystwyty University, Wales, U.K. in 2009. He was a research associate under the supervision of his PhD supervisor, Professor Mark H. Lee, at the Aberystwyth University from 2009 to 2010. He is currently an Associate Professor with the Cognitive Science Department, Xiamen University, China.
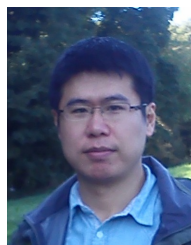

Zheming Zuo received the B.Eng. degree in Computer Science and Technology from the Southeast University, Nanjing, China, and the M.Sc. degree in Computer Science from the University of Birmingham, Birmingham, U.K., in 2012 and 2014, respectively. He is currently a Ph.D. student at Northumbria University with research interests including computer vision, machine learning, and fuzzy logic systems.

Ran Song is currently a Professor with the School of Control Science and Engineering at Shandong University, China. He graduated from University of York in 2009 with a Ph.D. degree. From 2009 to 2013, he was a postdoctoral research associate at Aberystwyth University. He moved to University of Brighton in 2014 as a senior lecturer. He has published more than 40 papers in some top-ranked international journals and conference proceedings. His research interests lie in 3D shape analysis and deep learning.

Longzhi Yang (M'12-SM'17) is the Director of Learning and Teaching and an Associate Professor in the Department of Computer and Information Sciences at Northumbria University, U.K. His research interests include computational intelligence, machine learning, big data, computer vision, intelligent control systems, robotics and the applications of such techniques under real-world uncertain environment. He is the founding Chair of the IEEE Special Interest Group on Big Data for Cyber Security and Privacy. He received the Best Student Paper Award at the 2010 IEEE International Conference on Fuzzy Systems. He is a senior member of IEEE, a professional member of British Computer Society, and Senior Fellow of Higher Education Academy of United Kingdom.

Yanpeng Qu (M'11) received a Ph.D. degree in Computational Mathematics from Dalian University of Technology, China. He is an Associate Professor with the Information Science and Technology College at Dalian Maritime University, China. His current research interests include granular computing, neural networks, pattern recognition, datamining, and real-world applications of such techniques for intelligent decision support.