Pattern Recognition 46 (2013) 1810–1818

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition

Li Liu, Ling Shao *, Peter Rockett

Department of Electronic and Electrical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK

ARTICLE INFO

Available online 12 October 2012

Keywords:
Action recognition
Pyramidal motion features
Boosted key-frame selection
Correlograms

ABSTRACT

In this paper we propose a novel method for human action recognition based on boosted key-frame selection and correlated pyramidal motion feature representations. Instead of using an unsupervised method to detect interest points, a Pyramidal Motion Feature (PMF), which combines optical flow with a biologically inspired feature, is extracted from each frame of a video sequence. The AdaBoost learning algorithm is then applied to select the most discriminative frames from a large feature pool. In this way, we obtain the top-ranked boosted frames of each video sequence as the key frames which carry the most representative motion information. Furthermore, we utilise the correlogram which focuses not only on probabilistic distributions within one frame but also on the temporal relationships of the action sequence. In the classification phase, a Support-Vector Machine (SVM) is adopted as the final classifier for human action recognition. To demonstrate generalizability, our method has been systematically tested on a variety of datasets and shown to be more effective and accurate for action recognition compared to the previous work. We obtain overall accuracies of: 95.5%, 93.7%, and 36.5% with our proposed method on the KTH, the multiview IXMAS and the challenging HMDB51 datasets, respectively.

1. Introduction

Human action recognition has attracted a great deal of attention due to its potential value and wide usage in a variety of areas, such as video search and retrieval, intelligent surveillance systems and human–computer interaction.

Typically, a scheme for human action recognition is based on either global or local feature extraction. Local feature methods usually follow the path of using an unsupervised technique, such as the method of Dollár et al. [1], to directly extract interest points from raw data followed by a 'bag of features' scheme to construct a code-book onto which raw features are mapped; histograms are normally used to represent each action. Finally, these representations are fed to a classifier. Methods based on local features, however, cannot usually achieve good results for human action recognition due to a dependence on the feature detector adopted. A good detector can extract more meaningful and significant features but identifying the best feature detector a priori is not generally possible. Also, it is obvious that just using a histogram representation of actions in a 'bag of features' scheme may lead to the loss of some discriminatory information in both the spatial and temporal dimensions, and influence the final recognition accuracies.

On the other hand, global feature methods consider the action as a volume in space/time and do not need feature detection—such methods are considered to be more holistic and informative for action recognition. Global features, however, can be sensitive to the background in the action sequence so that clutter, and indeed partial occlusion, may have a significant influence on action recognition accuracy.

In this paper, a new method based on key-frame selection is developed for human action recognition which identifies the frames carrying the most discriminative information. We extract a Pyramidal Motion Feature (PMF) for each frame of an action sequence but since not all the motion features are necessarily useful for action recognition, we use the AdaBoost learning algorithm [2] to select key frames for each action sequence. Each of these boosted key frames represents a typical motion pattern at one instant of the action sequence and the probabilistic distribution and temporal relationships of these frames are represented by a correlogram. Finally, a Support-Vector Machine (SVM) [3] is utilised as a classifier for recognizing actions. From a series of systematic experiments, we demonstrate that our method achieves results superior to the previously published work.

The main contributions of this paper are summarised as follows:

Firstly, a Pyramidal Motion Feature (PMF) is proposed to represent action sequences. We have applied an optical-flow algorithm combined with biologically inspired features to produce a new feature descriptor which is informative for action recognition.

* Corresponding author. Tel.: +44 114 222 5841.
E-mail address: ling.shao@sheffield.ac.uk (L. Shao).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.10.004

Secondly, we demonstrate that the efficacy of the AdaBoost learning algorithm for selecting key frames from each action video sequence. Unlike the general 'bag of features' approach, a correlogram is used to represent the co-occurrence probability of an action within the boosted key frames.

We compare the results of our method with the other previously published techniques. In terms of the final recognition accuracies, our methods are shown to be more accurate for human action recognition.

The remainder of this paper is organised as follows: In Section 2 we survey some related work. In Section 3 we describe details of our method; experimental results are given in Section 4. We draw conclusions in Section 5.

2. Related work

Much previous work on human action recognition has used frame representations. The silhouette representation, which records the pose of an action at a particular instant, was combined with a correlogram by Shao et al. [4] to achieve action classification. Wang et al. [5] have also applied an extended Histogram of Oriented Gradients (HOG) algorithm to represent frames for action recognition. Other frame representation approaches have been presented in [6–8]. Not all the frames in a sequence, however, are relevant to the action in question since some capture less meaningful information, or even describe a pose common to all action sequences, which may have a big influence on final classification performance.

Due to the above-mentioned drawbacks, another body of work has sought to select the most representative frames as *key* for action recognition instead of using the whole action sequence. Zhao and Elgammal [9] have proposed a method for human action recognition based on selecting key frames from a video sequence and representing them with the distribution of local motion features and their spatio-temporal arrangements. In their method, a small set of the most discriminative frames are selected by comparing their discriminative power for each independent action. A scheme of frame-by-frame voting is used for the action classification. Cooper and Foote [10] have presented a key-frame selection technique based on capturing similarity to the represented segment and preserving the differences with other segment key frames. In addition, Zhuang et al. [11] have used unsupervised clustering for key-frame selection. Cao et al. [12] have developed the novel approach of *key-pose* selection which utilises a PageRank-based centrality measure to select key poses for action recognition. Gong et al. [13] have also used a key-pose selection technique based on a local-motion energy optimization criterion to identify the frames with the most discriminatory pose motion information.

In this paper, we base our architecture for key-frame selection on the AdaBoost learning algorithm. The use of AdaBoost for feature selection in computer vision is fairly recent although it has mostly been applied to 2D data, principally for face recognition [14–16]. The work by Fathi and Mori [17] used mid-level motion features for action recognition by adopting boosted low-level optical flow information. Kellokumpu et al. [18] have also used AdaBoost to select the most discriminative features for human action recognition. Other researchers have applied an AdaBoost feature-selection scheme for real-time object detection [19], fast pedestrian recognition [20], and the retrieval of actions in movies [21] with remarkable improvements in recognition accuracy compared to the previous methods. We too apply the AdaBoost algorithm as a critical part of our methodology for selecting the most discriminative key frames.

3. Methodology

In the area of action recognition, attention is frequently focused on changes in the position of a subject with respect to time. Since motion information can reasonably describe action, we extract motion features by applying an optical-flow algorithm. In this paper, we extract Pyramidal Motion Features (PMFs) from the optical-flow information in each frame of an action sequence and use a supervised machine learning method (AdaBoost) to select the subset of frames with the most discriminatory motion features. A correlogram is then utilised to represent the action sequence instead of the more usual histogram representation—this correlogram representation is demonstrated to be more accurate for action recognition. Finally, we employ a Support-Vector Machine (SVM) to classify the actions. The overall structure of our method is shown in Fig. 1.

3.1. Sequence pre-processing and optical flow extraction

We pre-process all raw video sequences by determining 3D bounding boxes which localise the particular actions. For both spatial and temporal dimensions, these bounding boxes are scaled by linear interpolation to equal sizes for all types of actions.

In order to produce stable measures of the moving regions of the action sequences, we first subtracted adjacent frames. Thus subtracting frames n_{i+1} and n_i gives an estimate of the difference frame midway between n_{i+1} and n_i, $n_{i+(1/2)}$. Applying the Lucas-Kanade [22] algorithm to calculate the optical flow between adjacent difference frames $n_{i+(1/2)}$ and $n_{i-(1/2)}$ produces an estimate of the optical flow vector field F in frame n_i. (Experimentally, the initial differencing operation yielded a more stable flow field estimate since it reduces the effect of background clutter.) The optical flow vector field F is then split into horizontal and vertical components F_x and F_y. To reduce the effects of noise we have applied a spatial Gaussian blur ($\sigma_{x,y}=2$) to the magnitude images of F_x and F_y.

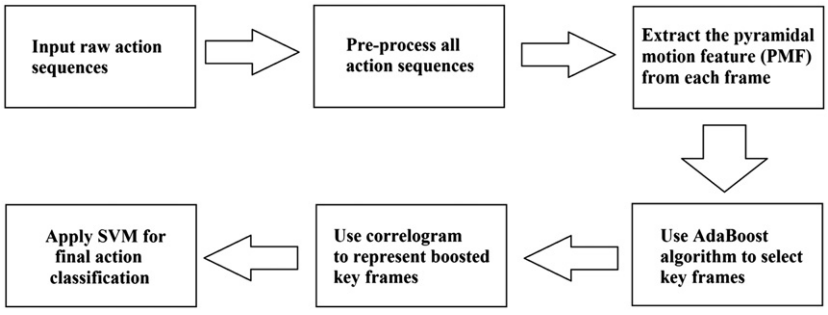

Fig. 1. The main structure of our proposed method.

3.2. Biologically inspired feature framework

Our biologically inspired feature extraction framework is based on the work of Olva and Torralba [23] which simulates the scene classification processes in the mammalian visual cortex. In computer vision, biologically inspired features are attractive for visual recognition since they encode intensity information while tolerating motion, translation, and scaling. The feature extraction process focuses on various orientations and intensity contrasts.

Our approach is motivated by mechanisms in the visual cortex which comprise four different layers, i.e. S1, C1, S2, and C2—see Hubel and Wiesel [24]. The C1 images mimic complex cells in the visual cortex and form an orientation feature map using the HMAX model [25]. We first calculate S1 feature maps by applying a Gabor convolution kernel at multiple scales and orientations to the target images. The Gabor filter function is as follows:

$$G(x,y) = \exp\left(-\frac{(X^2+\gamma^2Y^2)}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}x\right) \quad (1)$$

where $X = x\cos\theta - y\sin\theta$ and $Y = x\sin\theta + y\cos\theta$.

In our method, we define Gabor filters at eight different scales from 5×5 to 19×19 at size increments of 2 pixels. In addition, four different orientations (0–135° at 45 degrees increments) are adopted. In this way, $8\times4=32$ S1 feature maps are calculated. To obtain the C1 images, we apply the *max pooling* technique at various window sizes (from 6×6 to 12×12 with size increments of 2 pixels) to down-sample the adjacent scales of the S1 feature maps with the same orientation. In max pooling, we consider an $n\times n$ image patch in the S1 image and take as the corresponding pixel value of the n-fold down-sampled C1 image, the maximum value in the $n\times n$ S1 image patch. A similar multiple-orientation mechanism seems to be used in the human brain to perceive visual actions while maintaining invariance to motion, scale, and translation.

In addition to the C1 images, we construct *intensity* images also inspired by bioscience. Intensity images simulate the perception of nerve cells which typically respond to the regions of an input image where large changes of colour occur. Thus they focus on the intensity information which is commonly one of the most significant features for representing target objects in computer vision. Song and Tao [26] have also recently proposed such a scheme for scene classification.

In this paper, we build intensity images by adopting a centre-surround technique which has been widely used in the area of image analysis, especially for object edge enhancement and detection. We first apply a Gaussian filter with different sampling scales to our smoothed optical-flow images $|F_x|$ and $|F_y|$ to construct an image pyramid of decreasing size with increasing scale. In our experiments, we define seven different levels ($\sigma^2 = 5$, 7, 9, 11,…) in the pyramid. Intensity images are obtained by computing the differences between various images in the Gaussian pyramid as illustrated in Fig. 2. A *surround* image of a smaller scale is subtracted from a *centre* image to yield an intensity image; since the surround image has smaller dimension than the centre image, it is up-sampled appropriately using linear interpolation. We have considered the pairings of centre images 2 and 3, and the surround images 6 and 7, giving four intensity images (2-5, 2-6, 3-6, and 3-7).

In this way, we obtain C1 and intensity images from the smoothed $|F_x|$ and $|F_y|$ optical-flow images. The C1 and intensity images are recoded as a single vector by concatenating the data originating from both the F_x and F_y images. This vector thus comprises the pyramidal motion feature for this frame. Fig. 3 depicts the flow of the processing.

3.3. Adaboost key-frame selection

In this paper, each frame of an action sequence is represented by a Pyramidal Motion Feature (PMF) as described above. Since not all frames in a sequence are relevant to the corresponding action, we employ the AdaBoost learning algorithm to select key frames sufficiently discriminatory to be distinguished from others.

The AdaBoost learning algorithm is a widely used machine learning method which employs an ensemble of weak classifiers to construct a strong classifier for pattern classification. Given a training set $\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_N,y_N)\}$ where $\mathbf{x}_i$ is the feature vector, in our case the pyramidal motion feature vector (Section 3.2). $y_i \in \{+1,-1\}$ is the class label of the corresponding feature vector. AdaBoost initialises all training data with equal weight values, D_i. For each training iteration, the values of weights are updated depending on the learning results, i.e. the weights of incorrectly classified patterns are increased and, conversely, the weights of correctly classified patterns are decreased. Therefore, with increasing numbers of iterations, the classifier focuses more on incorrectly classified patterns. Finally, all the weak classifiers are assembled into one strong one. In our method, we adopt Classification And Regression Trees (CART) [27] as our base (weak) classifiers. We continue running the AdaBoost algorithm until all patterns in the training set are correctly classified. After the AdaBoost learning procedure, we select as our reduced set of selected features those patterns with the smallest weight values since this implies that they are relatively easy to classify and hence highly discriminatory.

The conventional AdaBoost learning algorithm is limited to two-class problems. To address the present multi-class problem, we use the 'one-against-the-rest' technique to extend binary classification to multiple classes. For instance, we first label all feature vectors from one type of action as the positive samples, and the remaining vectors belonging to all the other types of actions as the negative samples. Then, the AdaBoost algorithm is run to select the more discriminatory features for the positive class. We repeat this procedure for each type of action (class). In this way, the more discriminatory feature representations of individual frames in a given sequence are selected to represent each type of action. Key frame selection is illustrated in Fig. 4.

3.4. The correlogram of selected frames

The selected key frames are used as inputs to a 'bag of features' scheme although the histograms originally used with this approach only indicate the probability of feature distributions. Instead of using histograms, our method represents each action video with a correlogram which considers both the statistical distribution, and the relationship among all frames in the temporal dimension. A correlogram extends the K-bins of a 1D histogram representation (where K is the size of the codebook obtained by K-means clustering) to a more informative $K\times K$ matrix representation.

Since some spatial and temporal information may be lost by using a histogram representation during the traditional bag of words procedure, we take such omitted information into an account and represent it with a correlogram matrix which was first proposed by Huang et al. [28] for image indexing. Inspired by Huang's work, we form our pyramidal motion feature correlogram. Each element in our correlogram matrix is calculated as the co-occurrence probability of two frames taking place with a certain time offset of each other—see Eq. (2).

$$C(cluster_i,cluster_j;\Delta t) = \sum_{frame=1}^{frame\ no.-\Delta t} W(cluster_i,frame) \times W(cluster_j,frame+\Delta t) \quad (2)$$

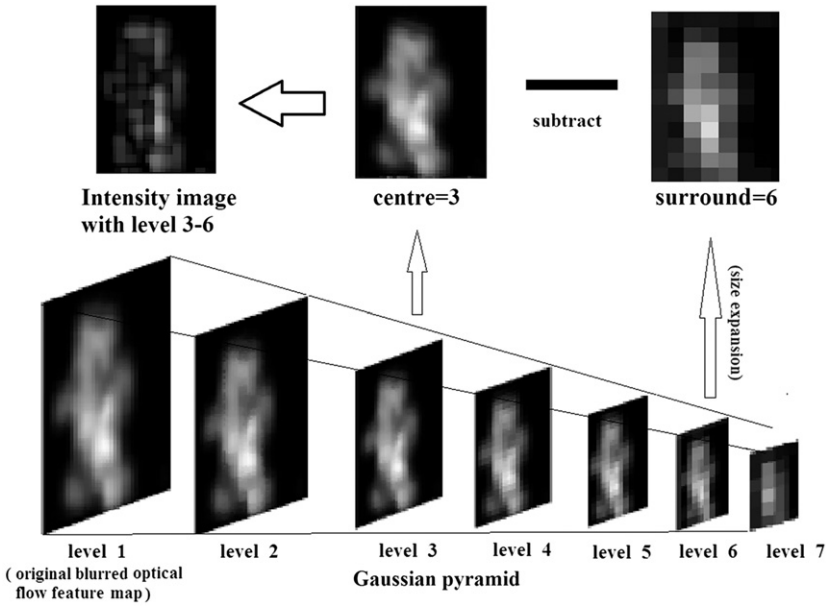

Fig. 2. Illustration of the formation intensity images.

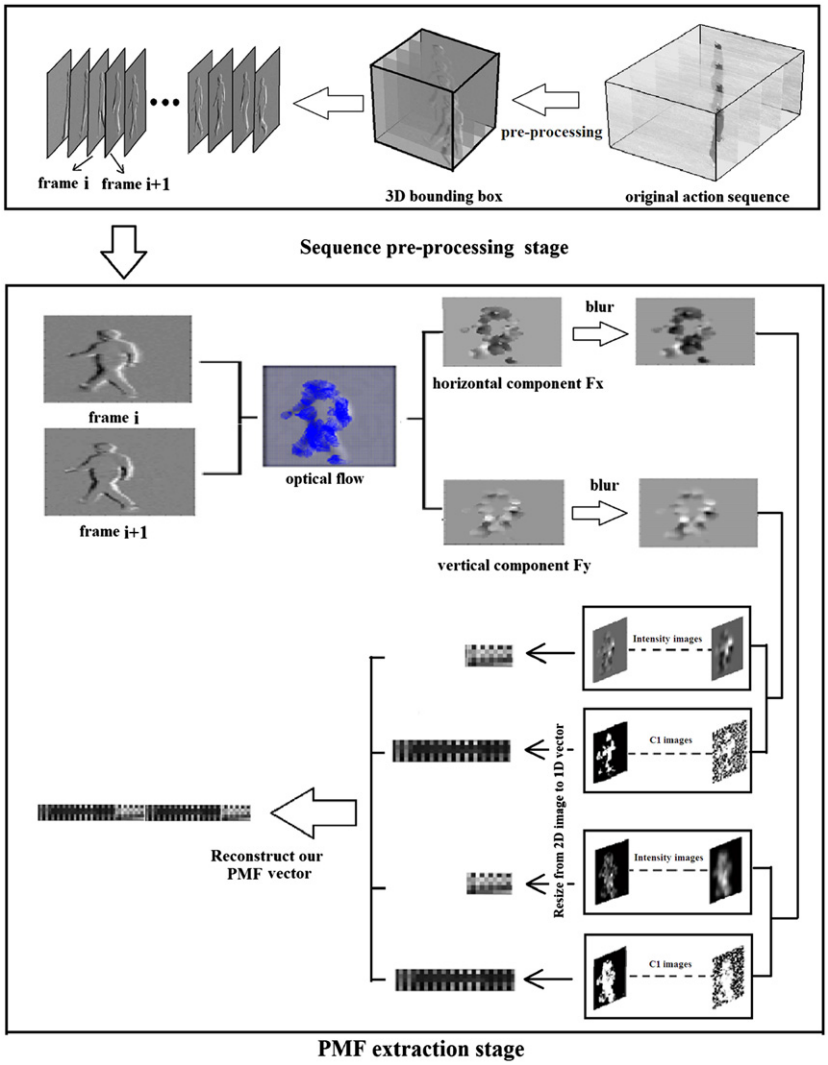

Fig. 3. The proposed feature extraction procedure: for one input action sequence, the Pyramidal Motion Feature (PMF) can be calculated from any two adjacent frames (i.e. frame i and frame $i+1$) and form a single representation vector.

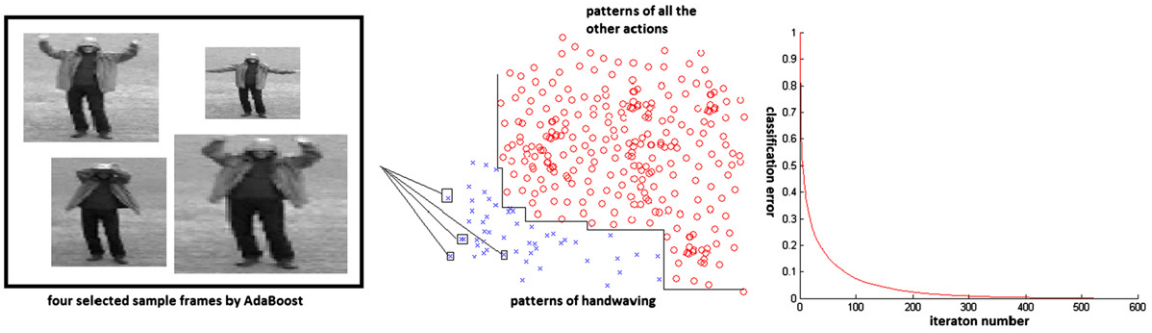

Fig. 4. The left sub-figure shows four boosted key-frame samples for the 'Handwaving' action. The middle sub-figure illustrates the final boundary for two kinds of patterns (e.g. one from 'Handwaving' and the other from all other action types); note that the key frames are remote from the decision surface. The right sub-figure shows the classification error with the increasing iterations during the AdaBoost learning.

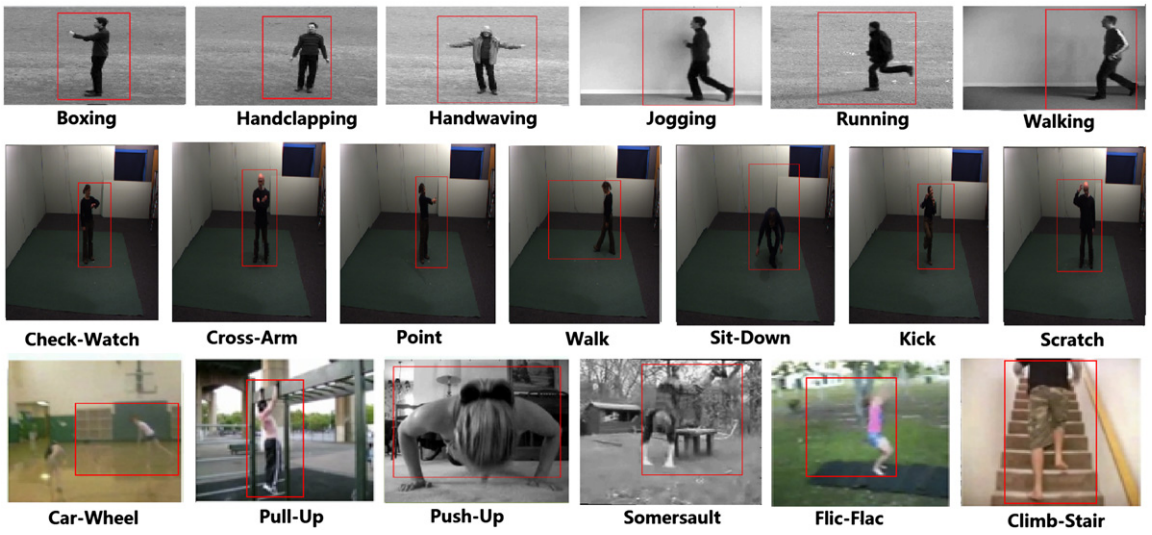

Fig. 5. Examples of the bounding boxes employed on typical frames from the three datasets. Row 1 = KTH dataset, row 2 = IXMAS dataset, and row 3 = HMDB51 dataset.

where $W(cluster_i,frame) = \exp(-\|centre-frame\|^2/2\sigma^2)$. Δt indicates the time offset. *Frame* and *centre* denote the pyramidal motion feature vector of one particular frame and the centre of a certain cluster (i.e. a visual word produced by *K*-means clustering). We use the Euclidean distance between the *frame* and *centre*.

In this paper, we use $K=120$ and to make the correlogram of a key-frame representation more distinctive, we construct it for three different time lags ($\Delta t = 1,2,4$). This correlogram matrix is reordered into a single vector to facilitate later classification using a SVM. The dimensions of this single vector are also reduced by supervised Linear Discriminant Analysis (LDA) prior to classification to remove redundancy. The details of the correlogram representation for action recognition can be found in [29].

4. Results

We have systematically evaluated our method using three different datasets: KTH [30], IXMAS [31], and HMDB51 [32] to demonstrate generalizability, and have compared our results with the other previously published reports.

Fig. 5 shows examples of the bounding boxes superimposed on typical frames from the three datasets. It is easy to see that these bounding boxes are rather coarsely determined, much larger than the human figures performing the actions, and do not necessarily locate the subject in the centre of the bounding box. Our approach is, therefore, fairly robust to determination of the bounding box.

4.1. KTH dataset

The KTH dataset contains six types of human action examples (i.e. boxing, hand clapping, hand waving, jogging, running, and walking) featuring 25 different subjects. Each action is performed in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors.

Following the pre-processing step mentioned in [33], the coarse 3D bounding boxes were extracted from all the raw action sequences and further normalised to the equal size of $100\times 100\times 60$. The Pyramidal Motion Feature (PMF) was then extracted from each frame of the pre-processed action sequences and a total of 34,800 features were obtained. We ran the AdaBoost selection procedure six times, once per action type, to select the top 10 discriminatory frames for each action video sequence. Fig. 6 shows the discrimination levels of frames in the KTH dataset. We can observe that different actions have different sets of key discriminative frames.

As is customary with this dataset, we performed "leave-one-out" cross validation over each of the 25 subjects (leave-one-person-out) to assess the accuracy of action recognition. The average accuracy is 95.5% on the KTH dataset and the corresponding confusion matrix is shown in Fig. 7. It is clear that good class separation has been obtained for all classes, the greatest confusion being between jogging and running which are two actions that would intuitively seem hard to reliably differentiate.

Since KTH is one of the most widely used datasets for human action recognition, the present and some previously published

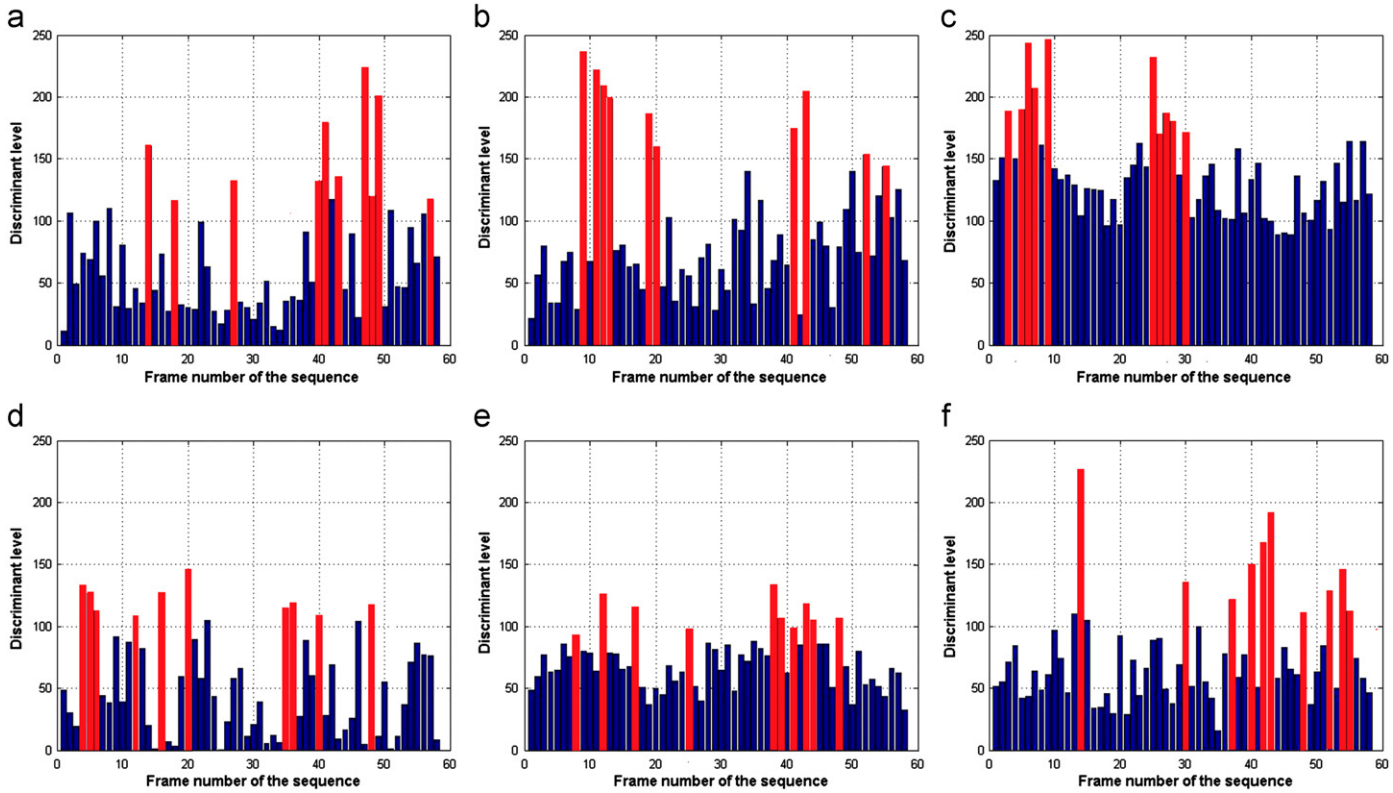

Fig. 6. (a) to (f) show the discrimination levels of sequence frames for different actions in the KTH dataset calculated from the corresponding weight values after Adaboost selection. (Here we plot discrimination level $= 100\ e^{-(weight)^2}$ for convenience of display.) The frames with the highest discrimination levels are selected as key frames. The red bars in each sub-figure indicate the top ten key frames selected for each action type. (a) Boxing, (b) Handclapping, (c) Handwaving, (d) Jogging, (e) Running, (f) Walking.

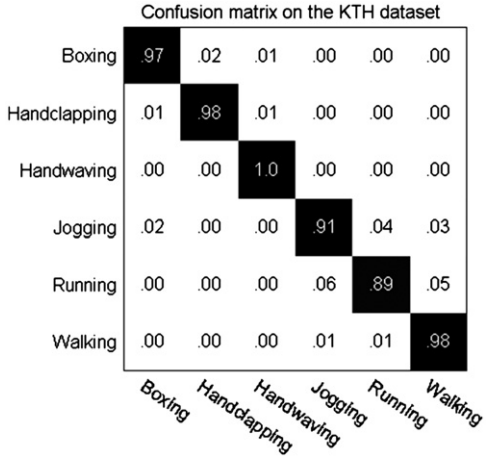

Fig. 7. The confusion matrix of final classification results for the KTH dataset.

results are compared in Table 1. The present method outperforms all the previously published results on this dataset.

4.2. IXMAS dataset

The Inria Xmas Motion Acquisition Sequences (IXMAS) motion dataset[1] is composed of 11 daily human actions performed by 10 actors and recorded from five different viewpoints. This is a more challenging dataset than KTH because actors can choose various positions or orientations in which to perform actions.

[1] http://4drepository.inrialpes.fr/public/viewgroup/6.

We extracted the bounding boxes by using foreground masks which are provided with the original dataset and then normalised them to the size of $100 \times 100 \times 75$. We have further extracted the Pyramidal Motion Feature (PMF) for each frame. In the key-frame selection phase, the top 10 discriminatory frames were selected using AdaBoost for each action sequence. The corresponding correlogram was computed using the selected features and a SVM employed for classification. Using "leave-one-out" cross validation, we have tested the performance of our method on each single-view camera as well as for the fused multiple-view cameras (i.e. using all the action sequences recorded by the five cameras together, and depicted in Fig. 8(a)). The overall camera fusion recognition accuracy reaches 93.7%, and the confusion matrix for the multiple-view cameras results is shown in Fig. 8(b). For comparison, Table 2 also compares our recognition results with those published in the other reports from which it is clear that our method outperforms all others, even the result of Weinland et al. [38] in which 3D reconstruction was applied before action recognition.

4.3. HMDB51 dataset

We have also experimented with the HMDB51 action-recognition dataset which collects action data from a variety of existing movies and online videos. The HMDB51 dataset contains 6849 clips divided into 51 action categories. Each category consists of at least 101 action clips. In our experiments, bounding boxes have been extracted from all the sequences using the masks released with dataset and initialised to the size of $250 \times 300 \times 120$. Due to the high computational costs, we ran the AdaBoost algorithm to select only the top 25 frames from each action clip. We followed the evaluation approach of Kuehne et al. [32] and split our data into three groups over which the average

Table 1
Comparison of action recognition accuracies in percentage (%) on the KTH dataset for different methods.

Methods	Actions						Average
	Boxing	Handclapping	Handwaving	Jogging	Running	Walking	
Our method	97	98	100	91	89	98	95.5
Dollár et al. [1]	93	77	85	57	85	90	81.2
Niebles et al. [34]	98	86	93	53	88	82	83.3
Ji et al. [35]	90	94	97	84	79	97	90.2
Jhuang et al. [36]	92	98	92	85	87	96	91.7
Liu and Shah [37]	98	95	96	89	87	100	94.2

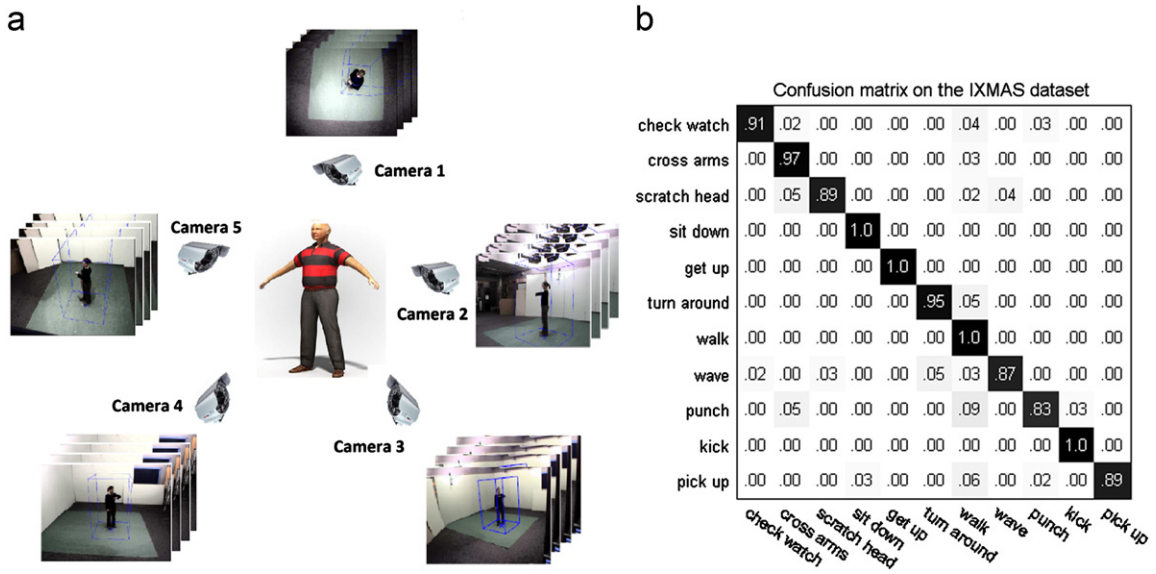

Fig. 8. IXMAS dataset. (a) Multi-view sketch map on the IXMAS dataset. (b) The confusion matrix of the multiple camera fusion result.

Table 2
Classification accuracies (%) of different methods for both single and multiple camera view on the IXMAS dataset.

Methods	Camera view					
	Cam1	Cam2	Cam3	Cam4	Cam5	Cam1–5
Our method	84.7	89.0	85.6	84.5	80.1	93.7
Varma and Babu [39]	76.4	74.5	73.6	71.8	60.4	81.3
Liu and Shah [37]	76.7	73.3	72.1	73.1	–	82.8
Wu et al. [40]	81.9	80.1	77.1	77.6	73.4	88.2
Weinland et al. [38]	–	–	–	–	–	93.3

accuracy is 36.5% for 3-fold cross-validation; the corresponding results are illustrated in Table 3. As far as we are aware, this is the first report of action recognition with the HMDB51 dataset. We only compare our results with the original paper [32] and our method achieves an obvious improvement on this dataset.

4.4. Summary of results

For comparison, we summarise the performance of four different recognition methods: our proposed method (PMF+AdaBoost+Correlogram+SVM), the same procedure without the Adaboost selection scheme (PMF+Correlogram+SVM), a scheme utilising a histogram to represent actions for classification instead of a correlogram (PMF+Histogram+SVM), and direct use of AdaBoost for selection and classification (PMF+AdaBoost). Comparative results are shown in Table 4 from which it is clear that the methods applying Adaboost selection and the correlogram representations achieve the highest recognition accuracy on the KTH, IXMAS, and HMDB51 datasets.

Table 3
Classification accuracies (%) on the HMDB51 dataset.

Methods	Splits			Average
	Split 1	Split 2	Split 3	
Our method	38.3	40.8	30.4	36.5
Kuehne et al. [32]	–	–	–	23.18

Table 4
Comparison of recognition performance (%) with/without the AdaBoost algorithm on the KTH, IXMAS, and HMDB51 datasets.

Methods	Dataset		
	KTH	IXMAS	HMDB51
PMF+Histogram+SVM	82.7	76.3	28.3
PMF+Correlogram+SVM	84.6	81.1	31.6
PMF+AdaBoost	91.7	86.8	33.8
PMF+AdaBoost+Correlogram+SVM	95.5	93.7	36.5

5. Conclusions

In this paper, we have employed the AdaBoost learning algorithm to select the most discriminative frames for a human

action recognition task. Instead of using hand-crafted interest point detectors and the original 'bag of features' approach, here we have utilised a supervised algorithm (i.e. AdaBoost) to select the subset of key, most discriminatory frames which are described by a Pyramidal Motion Feature (PMF). A correlogram is then used to form representations of each action sequence which is finally classified by a Support-Vector Machine (SVM). We have demonstrated that our boosted key-frame selection scheme produces an improvement in action recognition performance on three datasets: KTH, IXMAS, and HMDB51.

References

[1] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 2005, pp. 65–72.

[2] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: 2nd European Conference on Computational Learning Theory (EuroCOLT'95), 1995, pp. 23–37.

[3] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, in: IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997, pp. 130–136.

[4] L. Shao, D. Wu, X. Chen, Action recognition using correlogram of body poses and spectral regression, in: 18th IEEE International Conference on Image Processing, Brussels, Belgium, 2011, pp. 209–212.

[5] Y. Wang, G. Mori, Human action recognition by semilatent topic models, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1762–1774.

[6] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Computer Vision and Image Understanding 115 (2) (2011) 224–241.

[7] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: 9th IEEE International Conference on Computer Vision, vol. 2, Berkeley, CA, 2003, pp. 726–733.

[8] S. Carlsson, J. Sullivan, Action recognition by shape matching to key frames, in: IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision, vol. 1, Kauai, Hawaii, 2001, pp. 255–262.

[9] Z. Zhao, A. Elgammal, Information theoretic key frame selection for action recognition, in: British Machine Vision Conference (BMVC 2008), Leeds, UK, 2008, pp. 95–104.

[10] M. Cooper, J. Foote, Discriminative techniques for keyframe selection, in: IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 2005, p. 4.

[11] Y. Zhuang, Y. Rui, T. Huang, S. Mehrotra, Adaptive key frame extraction using unsupervised clustering, in: International Conference on Image Processing, vol. 1, Chicago, IL, 1998, pp. 866–870.

[12] X. Cao, B. Ning, P. Yan, X. Li, Selecting key poses on manifold for pairwise action recognition, IEEE Transactions on Industrial Informatics 8 (1) (2011) 168–177.

[13] W. Gong, A. Bagdanov, F. Roca, J. González, Automatic key pose selection for 3D human action recognition, in: 6th International Conference on Articulated Motion and Deformable Objects (AMDO'10), 2010, pp. 290–299.

[14] M. Zhou, H. Wei, Face verification using Gabor wavelets and AdaBoost, in: International Conference on Pattern Recognition, vol. 1, Hong Kong, 2006, pp. 404–407.

[15] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image and Vision Computing 27 (6) (2009) 803–816.

[16] S. Piyanuch, K. Deepak, H. Allen, Feature selection using AdaBoost for face expression recognition, in: 4th IASTED International Conference on Visualization, Imaging, and Image Processing, Marbella, Spain, 2004, pp. 261–286.

[17] A. Fathi, G. Mori, Action recognition by learning mid-level motion features, in: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1–8.

[18] V. Kellokumpu, G. Zhao, M. Pietikäinen, Dynamic textures for human movement recognition, in: ACM International Conference on Image and Video Retrieval, Xi'an, China, 2010, pp. 470–476.

[19] A. Treptow, A. Zell, Combining AdaBoost learning and evolutionary search to select features for real-time object detection, in: Congress on Evolutionary Computation, vol. 2, Portland, OR, 2004, pp. 2107–2113.

[20] S. Paisitkriangkrai, C. Shen, J. Zhang, Fast pedestrian detection using a cascade of boosted covariance features, IEEE Transactions on Circuits and Systems for Video Technology 18 (8) (2008) 1140–1151.

[21] I. Laptev, P. Pérez, Retrieving actions in movies, in: International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[22] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: DARPA Image Understanding Workshop, 1981, pp. 121–130.

[23] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision 42 (3) (2001) 145–175.

[24] D. Hubel, T. Wiesel, Receptive fields and functional architecture of monkey striate cortex, Journal of Physiology 195 (1) (1968) 215–243.

[25] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature Neuroscience 2 (11) (1999) 1019–1025.

[26] D. Song, D. Tao, Biologically inspired feature manifold for scene classification, IEEE Transactions on Image Processing 19 (1) (2010) 174–184.

[27] G. De'ath, K.E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology 81 (11) (2000) 3178–3192.

[28] J. Huang, S.R. Kumar, W.J. Mitra, M. Zhu, R. Zabih, Image indexing using color correlograms, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997, pp. 762–768.

[29] D. Wu, L. Shao, Silhouette analysis based action recognition via exploiting human poses, IEEE Transactions on Circuits and Systems for Video Technology (2012). http://dx.doi.org/10.1109/TCSVT.2012.2203731.

[30] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: International Conference on Pattern Recognition, vol. 3, Cambridge, UK, 2004, pp. 32–36.

[31] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3D exemplars, in: 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–7.

[32] H. Kuehne, H.J.E.G.T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: 13th IEEE International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 2556–2563.

[33] A. Yao, J. Gall, L.V. Gool, A Hough transform-based voting framework for action recognition, in: International Conference on Computer Vision and Pattern Recognition, 2010, pp. 2061–2068.

[34] J.C. Niebles, H. Wang, F. Li, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision 79 (3) (2008) 299–318.

[35] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, in: International Conference on Machine Learning, Haifa, Israel, 2010.

[36] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[37] J. Liu, M. Shah, Learning human actions via information maximization, in: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1–8.

[38] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2–3) (2006) 249–257.

[39] M. Varma, B.R. Babu, More generality in efficient multiple kernel learning, in: 26th Annual International Conference on Machine Learning, New York, NY, 2009, pp. 1065–1072.

[40] X. Wu, D. Xu, L. Duan, J. Luo, Action recognition using context and appearance distribution features, in: IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2011, pp. 489–496.

Li Liu received the B.Eng. degree in Electronic Information Engineering from Xi'an Jiaotong University, Xi'an, China, in 2011. He is currently working toward the Ph.D. degree in the Department of Electronic and Electrical Engineering, the University of Sheffield, UK. His research interests include human action recognition, scene and object classification, and genetic programming for visual feature extraction, representation and description.

Ling Shao received the B.Eng. degree in Electronic Engineering from the University of Science and Technology of China (USTC), the M.Sc. degree in Medical Image Analysis and the Ph.D. (D.Phil.) degree in Computer Vision at the Robotics Research Group from the University of Oxford.

Dr. Ling Shao is currently a Senior Lecturer (Associate Professor) in the Department of Electronic and Electrical Engineering at the University of Sheffield. Before joining Sheffield University, he worked for four years as a Senior Scientist in Philips Research, The Netherlands. His research interests include computer vision, pattern recognition, and video processing. He has published over 70 academic papers in refereed journals and conference proceedings and over 10 awarded patents and patent applications. Ling Shao is an Associate Editor of IEEE Transactions on Systems, Man and Cybernetics—Part B: Cybernetics, the International Journal of Image and Graphics, the EURASIP Journal on Advances in Signal Processing, and Neurocomputing, and has edited several special issues for journals of IEEE, Elsevier, and Springer. He has organized several workshops with top conferences, such as ICCV, ACM Multimedia, and ECCV. He has been serving as Program Committee member for many international conferences, including CVPR, ECCV, ICIP, ICASSP, ICME, ICMR, ACM MM, CIVR, BMVC, etc. He is a Fellow of the British Computer Society and a Senior Member of the IEEE.

Peter Rockett was born in London, England. He obtained a B.Sc. in Electronic Engineering in 1976 followed by an M.Sc. in Solid-state Electronics in 1977 and a Ph.D. in Semiconductor Physics in 1980, all from the University of Manchester Institute of Science and Technology (UMIST).

After holding various fellowships and research positions, he was appointed to a tenured faculty position in Electronic Systems in the Department of Electronic and Electrical Engineering, University of Sheffield, England, in 1990. His current research interests are in image processing, statistical pattern recognition, and multiobjective genetic programming.