Human Action Classification Using SVM_2K Classifier on Motion Features

Hongying Meng, Nick Pears, and Chris Bailey

Department of Computer Science, The University of York, York YO10 5DD, UK
{hongying, nep, chrisb}@cs.york.ac.uk

Abstract. In this paper, we study the human action classification problem based on motion features directly extracted from video. In order to implement a fast classification system, we select simple features that can be obtained from non-intensive computation. We also introduce the new SVM_2K classifier that can achieve improved performance over a standard SVM by combining two types of motion feature vector together. After learning, classification can be implemented very quickly because SVM_2K is a linear classifier. Experimental results demonstrate the method to be efficient and may be used in real-time human action classification systems.

1 Introduction

Digital video now plays an important role in entertainment, education, and other multimedia applications. It has become increasingly important to develop mechanisms that process, model, represent, summarize, analyse and organize the digital video information so that useful knowledge can be categorized automatically. Event detection in video is becoming an increasingly important application for computer vision, particular in the context of activity classification [1].

Recognizing actions of human actors from digital video is a challenging topic in computer vision with many fundamental applications in video surveillance, video indexing and social sciences. Feature extraction is the basis to perform many different tasks with video such as video object detection, object tracking, object classification, etc.

Appearance-based models are based on the extraction of a 2D shape model directly from the images, to be classified (or matched) against a trained one. Motion-based models do not rely on static models of the person, but on people motion characteristics [2][3][4][5][6].Motion feature extraction and selection is one of the key parts in these kinds of human action recognition systems. Bobick and Davis [2] use motion-energy images (MEI) and motion-history images (MHI) to recognize many types of aerobics exercises. Ogata et al [3] proposed another efficient technique for human motion recognition based on motion history images and an eigenspace technique. Schuldt et al [4] construct video representations in terms of local space-time features and integrate such representations with Support Vector Machine (SVM) [7] [8] classification schemes for recognition.

B. Gunsel et al. (Eds.): MRCS 2006, LNCS 4105, pp. 458–465, 2006.

Ke et al [5] generalize the notion of 2D box features to 3D spatio-temporal volumetric features for event detection in video sequences. Weinland et al [6] introduce Motion History Volumes (MHV) as a free-viewpoint representation for human actions in the case of multiple calibrated, and background-subtracted, video cameras. More recently, Wong and Cipolla [9] [10]proposed a new method to recognise primitive movements based on the motion gradient orientation image directly from videos.

In this paper, we build a fast human action classification system based on the SVM_2K classifier [11] [12] and some simple motion features. In comparison with a standard SVM [7] [8], the SVM_2K classifier can efficiently combine two distinct motion features together and achieve better classification performance.

2 Human Action Classification System

A suitable classifier should be at the core of any classification system. In this paper, we proposed a fast human action classification system based on the SVM_2K classifier . There are three main reasons for us to choose SVM_2K. Firstly, SVM_2K is a linear classifier, which means that it will be very easy and simple to implement in terms of classification, although the learning (training) is not as simple. Secondly, SVM_2K can deal with very high dimensional feature vectors, which means that we can choose the feature vectors without practical dimension limits. Finally, in comparison with standard SVM approaches, SVM_2K can achieve better performance by efficiently combining two types of motion feature together.

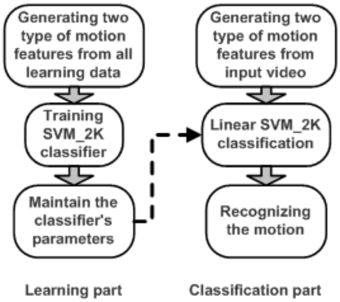
Generating two type of motion features from all learning data
Training SVM_2K classifier
Maintain the classifier's parameters
Generating two type of motion features from input video
Linear SVM_2K classification
Recognizing the motion
Learning part
Classification part

Fig. 1. SVM_2K based classification system. In learning part, two motion features were used to train SVM_2K classifier, and the obtained parameters were used in classification part.

A normal classification system includes two parts: a learning (or training) part and a classification part. These two parts of our classification system are showed separately in figure 1. The feature vectors are obtained using motion information directly from the input video. It is expected that the feature extraction algorithms should be as simple as possible. The high dimensional feature vector can be easily dealt with by SVM_2K classifier.

The learning part is processed using video data collected off-line. After that, the obtained parameters for the classifier can be used in a small, embedded computing device such as a FPGA (field-programmable gate array) or DSP (digital signal processor) based system, which can be embedded in the application and give real-time performance.

3 Motion Features

The recording of human actions usually needs large amounts of digital storage space and it is time consuming to browse the whole video to find the required information. It is also difficult to deal with this huge data in detection and recognition. Therefore, several motion features have been proposed to compact the whole motion sequence into one image to represent the motion. The most popular of these are the Motion History Image (MHI) and the Modified Motion History Image (MMHI). These two motion features have the same size as the frame of the video, but they maintain the motion information within them.

3.1 MHI

A motion history image (MHI) is a kind of temporal template. It is the weighted sum of past successive images and the weights decay as time lapses. Therefore, an MHI image contains past raw images within itself, where most recent image is brighter than past ones.

Normally, an MHI $H_\tau(u,v,k)$ at time k and location (u,v) is defined by the following equation 1:

$$H_\tau(u,v,k) = \{ \begin{matrix} \tau & if\ D(u,v,k) = 1 \\ max\{0, H_\tau(u,v,k) - 1\} & otherwise \end{matrix} \qquad (1)$$

where $D(u,v,k)$ is a binary image obtained from subtraction of frames, and τ is the maximum duration a motion is stored. An MHI pixel can have a range of values, whereas the Motion Energy Image (MEI) is its binary version. This can easily be computed by thresholding $H_\tau > 0$.

3.2 MMHI

Ogata et. al [3] use a multi-valued differential image to extract information about human posture because differential images encode human posture information more than a binary image such as a silhouette image. They propose a Modified Motion History Image (MMHI) defined as follows:

$$H_\delta(u,v,k) = max(f_l(u,v,k), \delta H_\delta(u,v,k-1)) \qquad (2)$$

where $f_l(u,v,k)$ is an input image (a multi-valued differential image), H_δ is the modified MHI, and parameter δ is a vanishing rate which is set at $0 < \delta \leq 1$. When $\delta = 1$, it was called as superposed motion image (SMI) which is the

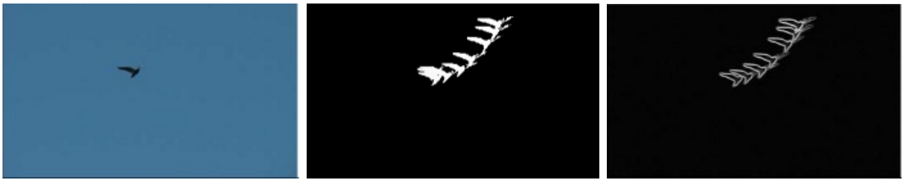

Fig. 2. In this video sample, a bird flys in the sky (left). The features MHI (middle) and MMHI (right) both have retained the motion information of the bird.

maximum value image generated from summing past successive images with an equal weight.

Figure 2 shows the motion features of MHI (b) and MMHI (c) of a bird flight in the sky (a). From these features, we can clearly determine how the bird flew in the sky even we didn't see the video clip, since these features retain the motion information within them.

4 SVM_2K Classifier

The two-view classifier SVM_2K is a linear binary classifier. In comparison with SVM, it can achieve better performance by combining two features together. The two-view classifier SVM_2K was firstly proposed in paper [11] in which the basic formation and fast algorithms were provided. It was shown that it worked very well in generic object recognition problems. Further theoretical study was provided in a later paper [12].

Suppose we have a data set $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, m\}$, where $\{\mathbf{x}_i\}$ are samples and have labels $\{y_i = \{-1, +1\}\}$ and we have two types of mapping ϕ^A and ϕ^B on $\{\mathbf{x}_i\}$ to get the feature vectors in two different feature spaces. Then SVM_2K classifier can be expressed as the following constraint optimization problem:

$$
\begin{array}{rc}
 & \min \frac{1}{2}(||\mathbf{w}_A||_2^2 + ||\mathbf{w}_B||_2^2) + \mathbf{1}^T(C^A\boldsymbol{\xi}^A + C^B\boldsymbol{\xi}^B + D\boldsymbol{\eta}) \\
 & \text{with respect to} \\
 & \mathbf{w}_A,\ \mathbf{w}_B,\ b_A,\ b_B,\ \boldsymbol{\xi}^A,\ \boldsymbol{\xi}^B,\ \boldsymbol{\eta} \\
 & \text{subject to} \\
Synthesis & \psi(\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A, \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle - b_B) \leq \eta_i + \epsilon, \\
subSVM1 & y_i(\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A) \geq 1 - \boldsymbol{\xi}_i^A, \\
subSVM2 & y_i(\langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle + b_B) \geq 1 - \boldsymbol{\xi}_i^B, \\
 & \boldsymbol{\xi}^A \geq 0,\ \boldsymbol{\xi}^B \geq 0,\ \eta \geq 0,\ i = 1, \ldots, m, \\
 & \boldsymbol{\xi}^A = (\xi_1^A, \ldots, \xi_m^A),\ \boldsymbol{\xi}^B = (\xi_1^B, \ldots, \xi_m^B), \\
 & \boldsymbol{\eta} = (\eta_1, \ldots, \eta_m).
\end{array}
\tag{3}
$$

In this formulation, $\mathbf{1}$ is a vector for which every component equals to 1. The constants C^A, C^B and D are penalty parameters. From this formulation, two SVM classifiers on feature (A) with parameters $(\mathbf{w}_A, \boldsymbol{\xi}^A,\ b_A)$ and feature (B) with parameters $(\mathbf{w}_B, \boldsymbol{\xi}^B,\ b_B)$ are combined together in one united form. ϵ is a small constant and $\boldsymbol{\eta}$ are associate slack variables. The important part of this formulation is the synthesis function ψ which links the two SVM subproblems by forcing them to be similar with respect to the values of the decision functions. As in paper [11], we use ψ as the absolute value of the differences for every $i = 1, \ldots, m$. That is,

$$\psi(\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A, \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle - b_B) = |\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A - \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle - b_B|.$$

In comparison with a standard SVM, this is a more complex constrained optimization problem and can be solved by quadratic programming by adding some constraints [11]. However, it is computationally expensive. Fortunately, an Augmented Lagrangian based algorithm was provided in [11] and this works very efficiently and quickly. SVM_2K can directly deal with the MHI and MMHI images as feature vectors with high deminsion in its learning and classification process. No segmentation or other oprations are needed.

5 Experimental Results

For the evaluation, we use a challenging human action recognition database, recorded by Christian Schuldt [4]. It contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4).

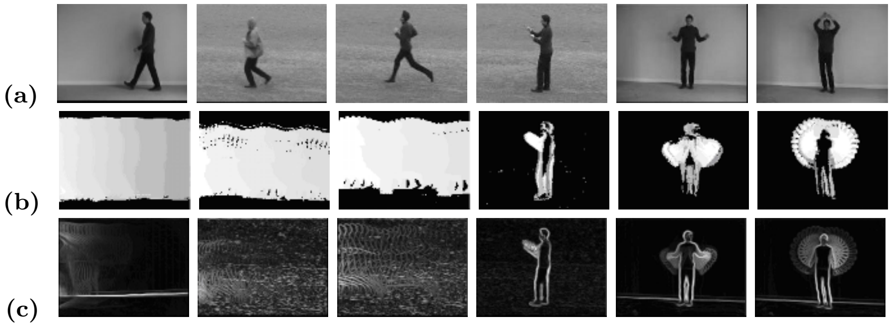

Fig. 3. Six types of human actions in the database: walking, jogging, running, boxing, handclapping and handwaving. Row (a) are the original videos, (b) and (c) are associate MHI and MMHI features.

This database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with 25Hz frame rate. The sequences were down-sampled to the spatial resolution of 160×120 pixels and have a length of four seconds in average. All sequences were divided with respect to the subjects into a training set (8 persons), a validation set (8 persons) and a test set (9 persons).

Figure 3 shows the examples in each type of human action and their associate MHI and MMHI motion features. In order to compare our results with the one in paper [5], we use same training set and testing dataset in our experiments. The only difference is that we didn't use the validation dataset in the learning.

Our experiments are carried out on all four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In the same manner as paper [5], each sequence is treated individually during the training and classification process. In all the following experiments, the parameters were chosen to be same. The threshold in differential frame computing was chosen as 25 and $\delta = 0.95$ in MMHI. The constants $C^A = C^B = D = 2$ and $\epsilon = 0.005$ in SVM_2K classification.

Table 1. Ke's confusion matrix [5], trace=377.8

	Walk	Jog	Run	Box	Clap	Wave
Walk	80.6	11.1	8.3	0.0	0.0	0.0
Jog	30.6	36.2	33.3	0.0	0.0	0.0
Run	2.8	25.0	44.4	0.0	27.8	0.0
Box	0.0	2.8	11.1	69.4	11.1	5.6
Clap	0.0	0.0	5.6	36.1	55.6	2.8
Wave	0.0	5.6	0.0	2.8	0.0	91.7

Table 2. SVM_2K's confusion matrix, trace=391.7

	Walk	Jog	Run	Box	Clap	Wave
Walk	68.1	21.5	9.7	0.0	0.0	0.7
Jog	27.1	50.0	20.8	1.4	0.0	0.7
Run	18.1	36.8	41.7	2.8	0.0	0.7
Box	0.0	0.0	0.0	100.0	0.0	0.0
Clap	0.0	0.0	0.0	34.0	60.4	5.6
Wave	0.0	0.0	0.0	22.2	6.3	71.5

Tables 1 show the classification confusion matrix based on the method proposed in paper [5]. Table 2 shows the confusion matrix obtained by our method based on two features MHI and MMHI. The confusion matrices show the motion label (vertical) versus the classification results (horizontal). Each cell (i, j) in the table shows the percentage of class i action being recognized as class j. Then trace of the matrices show the percentage of the correctly recognized action while the remaining cells show the percentage of misclassification. Note that our method obtained a better performance than Ke's method based on volumetric features. It should be mensioned here that in paper [4], the performance is slightly better where trace=430.3. But our system was trained as same as [5] to detect a single instance of each action within arbitrary sequences while Schuldt et al's system has the easier task of classifying each complete sequence(containing several repetitions of same action) into one of six classes.

From these tables, we can see that some actions such as boxing, hand clapping and handwaving are easy to recognise, while walking, jogging and running are difficult. The reason is that the latter three are very similar each other both from video sequences and the feature images.

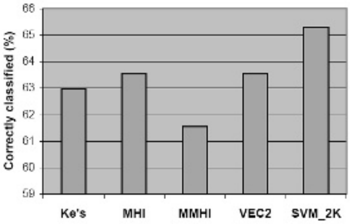

Fig. 4. Comparison results on the correctly classified rate based on different methods: Ke's method; SVM on MHI; SVM on MMHI; SVM on the concatenated feature (VEC2) of MHI and MMHI and SVM_2K on MHI and MMHI

In order to compare the performance of two classifier: SVM_2K and SVM only, a SVM was trained for each of the MHI and MMHI motion features separately and on the features (VEC2) created by concatenating them. The results are shown in the figure 4. It can be seen that SVM did well on MHI, but there is no improvement on VEC2. The SVM_2K classifier obtained the best results.

6 Conclusion

In this paper we proposed a new system for human action classification based on the SVM_2K classifier. In this system, we select the simple motion features MHI and MMHI. These features can retain the motion information of the actions and can be easily obtained with relatively low computational cost.

We introduced the classifier SVM_2K that can achieve better performance by combining two types of motion feature vectors MHI and MMHI together. SVM_2K can treat each MHI or MMHI image as a single feature vector where no segmentation or other oprations required on the features. After learning, fast classification for real-time applications can be implemented, because SVM_2K actually is a linear classifier.

In comparison with Ke's method, which is based on volume features, we use simple features and get better results. Experimental results also demonstrate that the SVM_2K classifier can obtain better results than a standard SVM on the same motion features.

If the learning part of the system is conducted off-line, this system has great potential for implementation in small, embedded computing devices, typically

FPGA or DSP based systems, which can be embedded in the application and give real-time performance.

Acknowledgements

This work is supported by DTI of UK and Broadcom Ltd.

References

1. Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. Computer Vision and Image Understanding **73** (1999) 428–440
2. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23** (2001) 257–267
3. T. Ogata, J.K.T., Ishikawa, S.: High-speed human motion recognition based on a motion history image and an eigenspace. IEICE Transactions on Information and Systems **E89** (2006) 281–289
4. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proc. Int. Conf. Pattern Recognition (ICPR'04), Cambridge, U.K (2004)
5. Y. Ke, R.S., Hebert., M.: Efficient visual event detection using volumetric features. In: Proceedings of International Conference on Computer Vision. (2005) 166–173 Beijing, China, Oct. 15-21, 2005.
6. Weinland, D., Ronfard, R., Boyer, E.: Motion history volumes for free viewpoint action recognition. In: IEEE International Workshop on modeling People and Human Interaction (PHI'05). (2005)
7. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines (and other kernel-based learning methods). Cambridge University Press, Cambridge, UK (2000)
8. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK (2004)
9. Wong, S.F., Cipolla, R.: Real-time adaptive hand motion recognition using a sparse bayesian classifier. In: ICCV-HCI. (2005) 170–179
10. Wong, S.F., Cipolla, R.: Real-time interpretation of hand motions using a sparse bayesian classifier on motion gradient orientation images. In: Proceedings of the British Machine Vision Conference. Volume 1., Oxford, UK (2005) 379–388
11. Meng, H., Shawe-Taylor, J., Szedmak, S., Farquhar, J.D.R.: Support vector machine to synthesise kernels. In: Deterministic and Statistical Methods in Machine Learning. (2004) 242–255
12. Farquhar, J.D.R., Hardoon, D.R., Meng, H., Shawe-Taylor, J., Szedmak, S.: Two view learning: SVM-2K, theory and practice. In: NIPS. (2005)