Discriminative Actions for Recognising Events

Karteek Alahari* and C. V. Jawahar

Centre for Visual Information Technology,
International Institute of Information Technology, Hyderabad 500032, INDIA.
jawahar@iiit.ac.in

Abstract. This paper presents an approach to identify the importance of different parts of a video sequence from the recognition point of view. It builds on the observations that: (1) events consist of more fundamental (or atomic) units, and (2) a discriminant-based approach is more appropriate for the recognition task, when compared to the standard modelling techniques, such as PCA, HMM, etc. We introduce *discriminative actions* which describe the usefulness of the fundamental units in distinguishing between events. We first extract actions to capture the fine characteristics of individual parts in the events. These actions are modelled and their usefulness in discriminating between events is estimated as a score. The score highlights the important parts (or actions) of the event from the recognition aspect. Applicability of the approach on different classes of events is demonstrated along with a statistical analysis.

1 Introduction

An event may be considered as a long-term temporally varying object, which typically spans over tens or hundreds of frames [1]. The problem of recognising events has received considerable research attention over the past few years [2–7]. It has gained importance because of its immediate applicability to surveillance, gesture recognition, sign language recognition, Human Computer Interaction, etc. [4, 8, 9]. Many approaches have been proposed in the past to recognise events. Early methods typically employed 2D or 3D tracking to temporally isolate the object performing the event. Subsequent to tracking, the event is recognised by extracting higher-order image features [6, 9]. An excellent review of such classical approaches for event recognition can be found in [2]. Owing to the inherent dynamism in events, Hidden Markov Models (HMMs) [10] and Finite State Machines [5] have been popular to address the event recognition problem. Furthermore, models such as HMMs provide elegant ways to incorporate the variability in a large collection of event data.

Another significant direction in event analysis research is to extract static image features from dynamic events [8, 11, 12]. Bobick and Davis [11] introduced Motion History and Motion Energy Images, which represent the recency and spatial density of motion respectively. In some sense their approach reduces the dimensionality of the event recognition problem from a 3D spatiotemporal space

* Currently at Oxford Brookes University, UK.

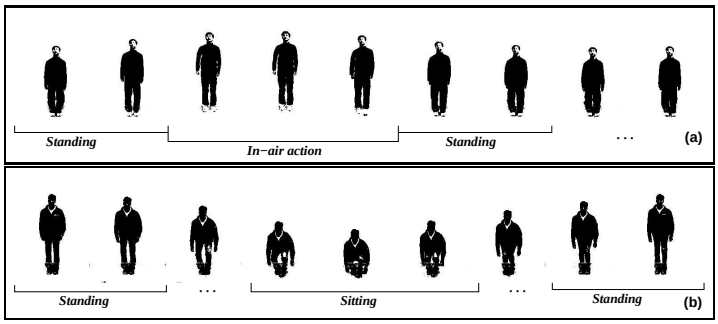

Fig. 1. A few sample processed frames (silhouettes) of the two events: (a) Jumping (first row), and (b) Squatting (second row), showing the constituent actions. Note the presence of a common action – *Standing* – between these events.

to a 2D image space. More recently Yilmaz and Shah [8] proposed a method to generate spatiotemporal volume (STV), in (x, y, t) space, using contours of the subject performing the event. The 3D "objects" are then recognised using differential geometric properties of STV. These methods either analyse the entire video sequence with a single image feature, which fails to capture the fine characteristics in events, or treat all parts of the sequence with equal importance, which leads to confusion in recognising highly similar events.

A method to recognise events using features which have optimal distinguishing characteristics is described in this paper. Our approach is motivated by the following observations.

- Events comprise of more fundamental (or atomic) units, which we refer to as *actions*. They are subsequences of the event sequences, and are a generalisation of the two extremes, namely the individual frames in an event (finest detail) and the entire video (coarsest detail). Analysing a video at the finest detail fails to capture the dynamism in events. On the other hand, analysing a video as a whole does not provide the fine details in various parts of the event sequence. *Actions* provide a natural mechanism to control the coarse-to-fine detail in the analysis.
- Due to the bulky nature of video sequences, it has been common to extract features in a low dimensional space. PCA is a popular modelling technique used to achieve this [2, 4]. However, it has been argued that discriminant techniques are more useful for the recognition task, when compared to modelling techniques [13].
- A direct discriminant analysis of video sequences, analogous to that performed on images, is not meaningful. This is because the relationships between parts of a video sequence are important, unlike the relationships between parts of an image. It is semantically useful to perform such an analysis at the *action* level.

Features, in the form of actions, are extracted to capture the fine characteristics of individual parts in the events. These actions are modelled and their useful-

ness in discriminating between events is estimated as a corresponding score. The score highlights the important parts (or actions) of the event from the recognition aspect. Using the estimated discriminatory scores and the corresponding action distances, a similarity measure is computed when comparing two events. The main advantages of our approach are as follows: (1) It requires minimal preprocessing of videos. In most situations where the video is recorded using a fixed camera, the background is relatively known, the silhouettes (see Fig. 1) can be extracted easily. (2) It is fast and does not require careful parameter tuning. (3) It is robust to the scheme used to extract actions from the events (Section 4.3).

1.1 Are events atomic?

Complex events such as people gesturing when interacting with others [14], playing Tennis [1], doing Aerobics [8], etc., are made up of more fundamental units. In fact, even simpler events such as a person squatting (see Fig. 1b) comprise of fundamental units: *standing* and *sitting*, in this case. Many researchers in the past have made similar observations [1, 3, 5, 15]. The fundamental (or atomic) units of events have been represented in many forms – as states of a stochastic finite automaton [5], components of PCA [15], the hidden states of HMMs [10], key frames in the event video [3], canonical poses [16], etc. There have also been approaches which analyse the event sequences in a window or block based fashion [1] to capture the granularity in the events.

All the above methods constitute a class of approaches which are designed to model the data in an optimal way. They deal with the representational aspects of events. It has been argued that such modelling techniques, suited for efficient representation, need not be the optimal for the classification task [13]. Discriminative models are more appropriate for the recognition task. However, it is not evident how these models can be derived in the context of video sequences.

This paper presents a discriminative model to recognise events effectively. In the past there have been a few attempts to use discriminant techniques for analysing video sequences, but are limited to either tracking [17] or gait-based human recognition [18]. We identify the actions in events, *i.e.* subsequences of the event sequences, which are more useful in discriminating between two events by analysing their statistical characteristics. The individual actions in the event are modelled to compute their discriminatory potential – the relative importance for distinguishing events – following a Fisher-like formulation [19]. To account for the statistical variability in each event, a collection of example event sequences is used. Each action together with its discriminatory potential is called a *discriminative action*. Using the discriminatory potentials (or weights) and the corresponding action distances for individual actions, a statistical distance measure is computed. Action distance denotes the similarity of two corresponding actions. In contrast, Han and Bhanu [18] use discriminant analysis only to extract features for human recognition based on gait. Also, unlike our approach the discriminative characteristics are not explicitly incorporated into the decision making process.

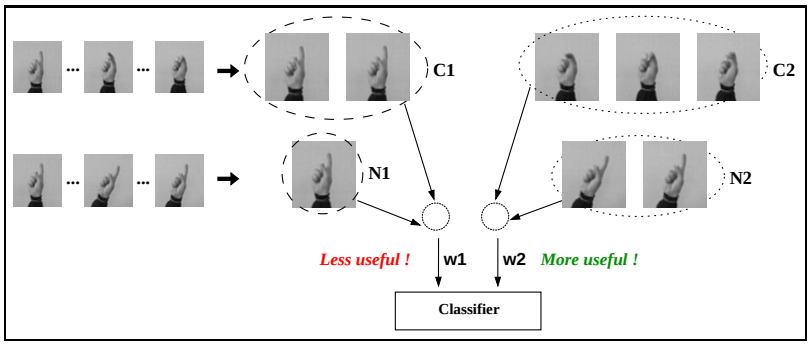

Fig. 2. Sample hand gesture frames showing two parts with different discriminatory potentials. Here the events *Click* (C) and *No* (N), the two events posses similar properties at the beginning of the sequences. The latter frames are more useful in the classification task when compared to the former frames. The individual segments (two shown here) of the video sequences are modelled and their discriminatory potential is combined to compute a similarity score.

The rest of this section discusses the motivation through an example. Section 2 presents the technical background along with an outline of the approach. The algorithm to obtain discriminant-based features for event sequences is given in Section 3. Section 4 presents results on two categories of event videos, namely hand gestures and human activities, along with a statistical analysis. Section 5 provides concluding remarks.

1.2 Motivation

To better appreciate the need for discriminative approaches for event recognition, consider the example illustrated in Fig. 2. It shows sample frames from two hand gesture [14] events: "*Click*", "*No*". The high degree of similarity among the gestures establishes the need to select the features which discriminate between the two event classes.

In the *Click* event (see Fig. 2) the subject moves his index finger vertically up and down, while in the *No* event the subject moves his index finger sideways horizontally, as if "saying" no to something. The two events appear to posses similar properties at the beginning of the sequences (where the finger remains in an almost stationary vertical state). As the complete event video sequence begins to appear over time, the distinguishing characteristics unfold, *i.e.* the latter frames of the sequence are more useful for discriminating between the two events when compared to the former frames. Hence, the latter frames should contribute more towards the decision making process. As shown in Fig. 2, the objective is to identify actions **C2** and **N2** which map to a feature space wherein the events are clearly distinguishable. The other parts (**C1** and **N1** in this example) owing to their similarity may not contribute much to the decision criteria. The popular pattern recognition approaches do not allow for such a scheme on video

data. They give equal importance to all the actions when comparing two event sequences, which may not be optimal, as in this case.

2 Semantic Discrimination of Events

Distinguishing between different parts of an event sequence, requires the need to "weigh" them appropriately when computing the decision criterion. This is in the spirit of Discriminant Analysis and Statistical Pattern Recognition techniques.

Fisher Discriminant Analysis (FDA) is a popular feature extraction scheme for 2-class problems [19]. It has been used to compute Fisherfaces in the image domain, which are optimal for recognition tasks [13]. FDA finds an optimal direction φ along which the between-class variance is maximised and the within-class variance is minimised. The criterion function $J(.)$ is defined as

$$J(\varphi) = \frac{\varphi^T \mathbf{S_b} \varphi}{\varphi^T \mathbf{S_w} \varphi}, \tag{1}$$

where $\mathbf{S_w}$ and $\mathbf{S_b}$ are the within-class and the between-class scatter matrices. The function $J(.)$ is maximised to compute the optimal φ for discriminating between the patterns. It is shown that any vector φ which maximises the Fisher criterion in Equation 1 satisfies $\mathbf{S_b}\varphi = \lambda \mathbf{S_w}\varphi$ for some constant λ [19]. This can be solved as an eigenvalue problem. Thus, the discriminant vector φ is given by the eigenvector corresponding to the largest eigenvalue of ${\mathbf{S_w}}^{-1}\mathbf{S_b}$. Extensions of Fisher Discriminant Analysis such as Multiple Discriminant Analysis, Kernel Discriminant Analysis, incremental LDA have also been used in computing discriminant features.

2.1 Event recognition using Discriminative Actions

Consider two video sequences $\mathbf{A}$ and $\mathbf{B}$ which belong to events (classes) $\mathcal{A}$ and $\mathcal{B}$ respectively. They represent a sequence of image frames where the corresponding event, like Click, No, etc., is captured.

The similarity between the two video sequences $\mathbf{A}$ and $\mathbf{B}$ can be computed by comparing the sequences directly. If the sequences are of different lengths, say due to variation in frame rate of video capture or duration of the event, a normalisation can be done by resampling. However, this naive comparison of video data frame-by-frame is not valid since the event of interest is macro in nature and cannot be captured from one sample frame. An appropriate intermediate subsequence is chosen for the representation to overcome this problem [1]. The problem we address is identification of the contribution of each of these subsequences (or actions) for the global dissimilarity/discriminative information for the given video sequences.

Let $\mathbf{A}^k$ and $\mathbf{B}^k, k = 1, 2, \ldots, s$ be the s actions extracted from the video sequences $\mathbf{A}$ and $\mathbf{B}$ respectively. Discriminative actions from a collection of event examples are computed as follows. Each action is represented as a corresponding static image by modelling its inherent dynamism. It is then modelled using

Discriminant History Images. This produces s images each for both the events. The discriminatory potentials computed for different parts of the video sequences and the action-action distance metrics are used to compute a weighted decision score when recognising a new video sequence.

Before segmenting the video sequences temporally, to extract the actions, they need to be aligned to account for the difference in frame capture rate. It is assumed that the collection of input video sequences is either already aligned or is captured at a uniform frame rate. A further discussion on the alignment schemes is beyond the scope of this paper. When the video sequences are captured at a uniform frame rate, the sequences are already aligned, and are directly segmented temporally. This process can also be understood as that of analysing the video sequence in a window-based fashion [1]. The number of actions is determined based on the set of events under consideration. For the experiments on videos sequences captured at 25 fps, with about 150 frames each, 6 actions are used, with the assumption that each action is performed in approximately 1 second.

3 Recognition based on Discriminative Actions

In this section, the technical details of the approach to identify *discriminative actions*, and subsequently use them to recognise events are presented. The separability of the two events is maximised and the variability within the event is minimised to compute these actions.

3.1 Computing Discriminative Actions

Representing Actions: Each action $\mathbf{A}^k$ consists of a set of image frames that describe the inherent dynamism in the action. The action characteristics are modelled using Motion History Images (MHI), which capture the dynamism in events, proposed by Bobick and Davis [11]. Although other modelling techniques are applicable in this context, for the results in this paper MHI features are used. They represent *how* motion is occurring in the actions. Given $N_{\mathcal{A}}$ and $N_{\mathcal{B}}$ instances each for the events $\mathcal{A}$ and $\mathcal{B}$, the MHI of the j th instance of the action $\mathbf{A}^k$ is denoted by $\theta^k_{\mathcal{A}j}$. Similarly, $\theta^k_{\mathcal{B}j}$ for the j th instance of the action $\mathbf{B}^k$. From [11], the intensities at pixels in the history image at time instant t, $H_\tau(t)$, are a function of the temporal history of the motion of the corresponding pixels. It is defined as $H_\tau(t) = \tau, \text{if I(t) = foreground}; \max(0, H_\tau(t-1)-1), \text{otherwise}$, where τ is a pre-determined constant and I(t) = foreground denotes the set of all pixels belonging to the event-performing subject. History Image features are computed for the last frame of every action. This provides exactly one History Image feature for each action. For instance, if there are p^k_j image frames in the j th instance of the action $\mathbf{A}^k$, $\theta^k_{\mathcal{A}j} = H_\tau(p^k_j)$. MHI features of a few sample video segments are illustrated in Fig. 3a. The motion trails of these actions clearly show how the motion is occurring. To enhance the discriminating characteristics between the two events, the relevance of individual actions for the recognition task is computed.

Computing the Discriminatory potential: The usefulness of a k th action for the recognition task is identified by $\varphi_k, k = 1, 2, \ldots, s$. It is computed such that the action features have optimal distinguishing characteristics along the direction of the vector φ. The within-class scatter (variability within events) is minimised and the between-class scatter (separability of events) is maximised for this. These scatter matrices are defined as

$$\mathbf{S_w} = \sum_{i \in \{\mathcal{A},\mathcal{B}\}} \sum_{j=1}^{N_i} (\theta_{ij} - \bar{\theta}_i)(\theta_{ij} - \bar{\theta}_i)^T,$$
$$\mathbf{S_b} = (\bar{\theta}_{\mathcal{A}} - \bar{\theta}_{\mathcal{B}})(\bar{\theta}_{\mathcal{A}} - \bar{\theta}_{\mathcal{B}})^T,$$

where the number of instances in class i is denoted by N_i, the symbols without the superscript k denote the sequence features with the action representations (MHIs) computed for each action stacked as rows, and the mean over the instances of a class i is given by $\bar{\theta}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \theta_{ij}$. Also, $(\theta_{ij} - \bar{\theta}_i)$ is the distance measure defined in the representation space. Here, the $s \times s$ matrices $\mathbf{S_w}$ and $\mathbf{S_b}$ capture the within-class and between-class scatters at the action level. Each entry of $\mathbf{S_b} = \{b_{ij}\}$ represents the variance between actions $\mathbf{A}^i$ and $\mathbf{B}^j$ over the set of all instances. Maximising the objective function in Equation 1 results in a discriminant vector of length s along which the classes possess large discriminating characteristics. Fig. 3b shows the actions and their corresponding discriminatory potentials for the event pair Click vs No. Discriminative actions are computed from a collection of example event video sequences. This constitutes the training phase of the proposed approach which is summarised below.

1. Align all the event video sequences in the training set with respect to a template video sequence, and segment them temporally to obtain s subsequences (or actions) for all the instances in the two classes $\mathcal{A}$ and $\mathcal{B}$. If it is known that the instances are captured at a uniform rate, segment them temporally.
2. Use Motion History Images (MHI) to compute the action representations and obtain the features: $\{\theta^k_{\mathcal{A}j}, \theta^k_{\mathcal{B}j}\}_{k=1}^{s}$.
3. Compute the discriminant vector φ, whose elements denote the relative importance of each action, by minimising the objective function $J(.)$ according to Equation 1.

3.2 Recognising Events

Let $\mathbf{T}$ be the event sequence which is to be recognised. It is labelled as class i^* according to $i^* = \arg\min_{i \in \{\mathcal{A},\mathcal{B}\}} D_\varphi(\mathbf{T}, i)$, where $D_\varphi(\mathbf{T}, i)$ defines the cost of recognising the sequence $\mathbf{T}$ as the sequence i in the discriminative feature space. The matching cost $D_\varphi(\mathbf{T}, \mathcal{A})$ is given by $D_\varphi(\mathbf{T}, \mathcal{A}) = f(\varphi_1 \ldots \varphi_s, \theta^1_{\mathbf{T}} \ldots \theta^s_{\mathbf{T}}, \theta^1_{\mathcal{A}} \ldots \theta^s_{\mathcal{A}})$.

The MHI features $\theta^1_{\mathbf{T}}, \ldots, \theta^s_{\mathbf{T}}$ of the actions from the test sequence are computed as described before for the training set. The function $f(.)$ models D_φ as a combination of the action level matching costs $d^k(.)$ and the weights φ_k, which discriminate between the actions. In other words, $f(.) = \sum_{k=1}^{s} \varphi_k d^k(\theta^k_{\mathcal{T}}, \theta^k_{\mathcal{A}})$, where $d^k(.)$ is defined as the Euclidean distance between the two MHI feature vectors.

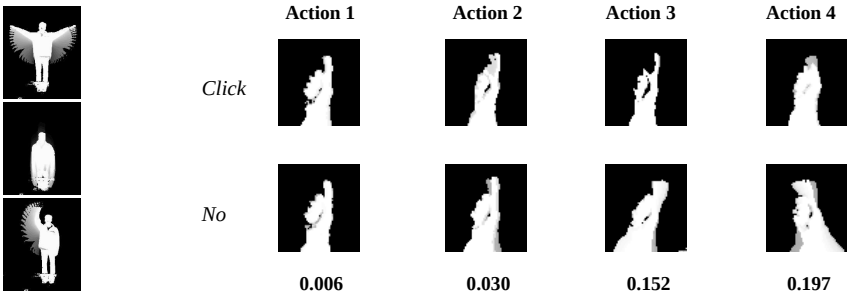

Fig. 3. (a) Motion History Images (MHI) of the events Flapping, Squatting and Waving respectively, clearly illustrating the motion trails. (b) MHI features computed using 4 actions of the events *Click*, *No*, and their corresponding discriminatory potential (shown in the last row). The first two actions have low discriminatory potential owing to their similarity. The last two actions are more useful for the classification task.

Significance of φ: Along the direction of the vector φ the ratio of between-class scatter and within-class scatter is maximised. When the data points, say, $\Theta_{\mathcal{T}} = [\theta^k_{\mathcal{T}}], k = 1, 2, \ldots, s$, and $\Theta_{\mathcal{A}} = [\theta^k_{\mathcal{A}}], k = 1, 2, \ldots, s$, are projected onto this direction as $\varphi^T \Theta_{\mathcal{T}}$ and $\varphi^T \Theta_{\mathcal{A}}$ respectively, each element of φ acts as a weight for the corresponding dimension. In this lower dimension space, the distance between two events $\mathcal{T}$ and $\mathcal{A}$ is expressed as a weighted linear combination of the distances along each dimension. The distance function $f(.)$ defined above can also be written as $f(.) = \varphi^T D(\Theta_{\mathcal{T}}, \Theta_{\mathcal{A}})$, where $D(\Theta_{\mathcal{T}}, \Theta_{\mathcal{A}}) = [d^1(\theta^1_{\mathcal{T}}, \theta^1_{\mathcal{A}}), \ldots, d^s(\theta^s_{\mathcal{T}}, \theta^s_{\mathcal{A}})]^T$. Assuming that the distance functions are metric, $f(.)$ can be simplified as $f(.) = D(\varphi^T \Theta_{\mathcal{T}}, \varphi^T \Theta_{\mathcal{A}}) = D_{\varphi}(\mathcal{T}, \mathcal{A})$. Thus, using metric distances the similarity between two events can be computed as a weighted linear combination of the action-level distances and the elements of the discriminant vector.

4 Experiments and Results

Results are presented on two classes of event video sequences – hand gestures and human activities. Both recorded and publicly available videos are used to test the applicability of the model.

4.1 Hand gestures

Recognising hand gestures has received a lot of attention in the recent past. It finds innumerable applications in HCI, Virtual Reality [4], wherein input to the computer can be regulated through various hand gestures, for instance controlling the visualisation of a CAD model. One of the challenges in hand gesture recognition is the high degree of similarity among the events. Hand gesture videos from Marcel's Dynamic Hand Gesture database [14] are used. It consists of 15 video sequences for each of the 4 dynamic hand gestures, namely Click,

No, StopGraspOk and Rotate. The data was divided into separate train and test sets. Results on three of the possible pairs – Click vs No, StopGraspOk vs Rotate, Rotate vs Click – which have a high degree of similarity between them are discussed. Sample frames of a couple of hand gestures are shown in Fig. 2. It can be observed that the two gestures are highly similar in the initial few frames and their distinguishing characteristics unfold over time in the latter frames. Following is the summary of the experiment conducted on this data set.

1. Minimal preprocessing (background subtraction and subsequent thresholding) is performed to the eliminate the background from the scene. The actions in the events are extracted according to the method discussed in Section 2.1.
2. Discriminative actions are estimated by modelling the actions as MHI features and computing their corresponding discriminatory potential, according to the method described in Section 3.1.
3. Given a new video sequence (of one of the trained categories) to recognise, we perform Step 1, model the actions as MHI features, and then use the estimated discriminant weights to compute the similarity score. The video sequence was labelled as discussed in Section 3.2.

The accuracy results on this data set are illustrated in Table 1. Results are compared to those obtained from a technique which gives equal importance to all parts of the sequence. No resubmission error is observed in the case where an optimally weighted distance measure is used. On average a percentage error reduction of 30.29 on about 200 video sequences is observed. Fig. 3b illustrates the Motion History Image features computed for 4 segments of *Click*, *No* gestures. It shows that the latter frames of the event sequence are more useful for the classification task.

4.2 Other activities

Recognition of events involving humans finds many applications in surveillance [2, 6, 11]. Most events performed by humans are marked by a considerable degree of commonality among them (for instance, see Fig. 1). This observation is exploited through the proposed discriminative action based method. For this experiment videos of 4 events, namely Jumping, Squatting, Limping, Walking, performed by 20 different people for an average duration of 6 seconds each, are used. These events occur with the subject either stationary or indulging in locomotion. In the former category, events Jumping and Squatting are considered, while in the latter category, Limping and Walking are considered. The videos are captured with a Panasonic Digital Video Camera at 24 fps. The data set is divided into distinct train and test sets. Minimal preprocessing is done on the video sequences as follows. In order to retain only the visually significant information, background subtraction and normalisation was done on all the frames. Motion compensation is also performed to centre the subject for the events where locomotion is involved. The events are temporally segmented into actions and are modelled using MHI features. The modelled actions are used to estimate the corresponding discriminatory potential. To recognise an unlabelled test event, the sequence

Event Pair	% Accuracy	
	Equal weights	Optimal weights
Click vs No	91	93
StopGraspOk vs Rotate	90	92
Rotate vs Click	87	92
Jumping vs Squatting	85	90
Limping vs Walking	87	91

Table 1. Recognition accuracy for about 200 video sequences. On an average a 30.29 percentage reduction in error is observed.

is preprocessed as above and the similarity measure is computed with respect to the two learnt event representations. The test video is then labelled as the event for which the weighted similarity measure is maximum (refer Section 3.2). The recognition accuracy results on these events are presented in Table 1. On an average, 32.05 percentage reduction in error is achieved.

4.3 Statistical Analysis

The proposed method improves the compactness and the separability of events. Within-event and between-event scatters in the standard and the discriminant-based feature spaces are computed to quantify the performance of the approach. This is done on a set of *Click* and *No* video sequences. Optimality of the feature space is defined in terms of the compactness (low variance within an event) and the separability (high variance between events) of the classes. Low within-event and high between-event scatters shown in Table 2, after transforming the features to a discriminant-based feature space, support our claim that this method identifies an optimal discriminant feature set.

The proposed approach is also not sensitive to the action extraction method used. It is observed that changing the action extraction method leads to negligible change in recognition accuracy. A noticeable change is observed only when the event is modelled as a single action. In the case of *Rotate* vs *Click* video pair, the average recognition accuracy was about 87% when modelled with a single action, and 91% when modelled with two actions. This is due to the fact that, the discriminatory potentials of different parts of the sequence are not exploited in a single action. Similar behaviour is observed on other video sequences.

The recognition scheme presented is applicable in a multiple class scenario as well. There are many ways of combining pairwise classifiers for solving multiple class problems. We use a Directed Acyclic Graph [19] is used to achieve this. The DAG is built following a one-vs-one architecture, where each node is a 2-class classifier. Multiple video sequences of 5 events are used to compute the recognition accuracy. All the 10 possible pairwise combinations of these events are trained to get the corresponding optimal weights. The results of this analysis are presented in Table 3. It shows the accuracy results in the multiclass scenario

Feature Space	Within-class scatter		Between-class scatter	
Standard	Class 1	5.025	Class 1 vs 2	6.174
	Class 2	4.619		
Discriminant-based	Class 1	3.907	Class 1 vs 2	15.958
	Class 2	2.794		

Table 2. Performance of the model in identifying an optimal discriminant feature space. The within-class and the between-class scatters for both the classes (*Click*: Class 1 and *No*: Class 2) in the standard and the proposed discriminant-based feature spaces are shown. The values are computed by considering the events to be comprising of 3 actions. Low within-event and high between-event scatter values indicate that our approach identifies a feature space wherein the classes are compact and well-separated.

and certain pairwise combinations of events. The discriminant weight approach shows significant improvement compared to the equal weight approach.

5 Conclusions

This paper addresses the issue of identifying the importance of different parts of a video sequence from the recognition point of view. It highlights the importance of feature selection for *recognising* rather than just *representing* events. An adaptive technique which chooses the important features from an event sequence is described. It demonstrates that a fixed feature selection scheme may not be appropriate for a wide class of events. This approach: (a) provides a mechanism to identify the video segments (actions) and their importance statistically, (b) is suitable for various domains involving analysis of sequential data such as video event sequences, online handwriting, etc., (c) is straight-forward to implement without requiring careful parameter-tuning, and (d) can be extended on the lines of Multiple Discriminant Analysis and Kernel Discriminant Analysis.

Acknowledgments. Karteek Alahari thanks the GE Foundation for financial support through the GE Foundation Scholar-Leaders Program 2003-2005.

References

1. Zelnik-Manor, L., Irani, M.: Event-Based Analysis of Video. In: Proc. CVPR. Volume II. (2001) 123–130
2. Gavrila, D.M.: The visual analysis of human movement: A survey. CVIU **73** (1999) 82–98
3. Sullivan, J., Carlsson, S.: Recognizing and Tracking Human Action. In: Proc. ECCV. Volume I. (2002) 629–644
4. Buxton, H.: Learning and understanding dynamic scene activity: a review. IVC **21** (2003) 125–136
5. Hongeng, S., Nevatia, R., Bremond, F.: Video-based event recognition: activity representation and probabilistic recognition methods. CVIU **96** (2004) 129–162

Multiclass classification			**Pairwise classification**		
Event	No. of misclassifications		Event Pair	No. of misclassifications	
	Equal weights	Optimal weights		Equal wts.	Optimal wts.
1	8/70	5/70	1 vs 2	12/140	5/140
2	6/70	6/70	2 vs 3	14/140	8/140
3	8/70	4/70	3 vs 4	9/140	6/140
4	5/70	4/70	4 vs 5	11/140	7/140
5	5/70	3/70	1 vs 5	10/140	6/140

Table 3. Event recognition results for 5 events in a multiclass scenario. The notation x/y denotes x misclassifications for y sequences. Some of the pairwise combination results (among the 10 possible combinations) are also shown. Similar results are observed on other pairs. In all cases the discriminant weight approach outperforms the equal weight approach.

6. Haritaoglu, I., Harwood, D., Davis, L.S.: W^4: Real-Time Surveillance of People and Their Activities. IEEE Trans. on PAMI **22** (2000) 809–830
7. Ryoo, M.S., Aggarwal, J.K.: Recog. of Composite Human Activities through Context-Free Grammar based Rep. In: CVPR. Volume 2. (2006) 1709–1718
8. Yilmaz, A., Shah, M.: Actions Sketch: A Novel Action Representation. In: Proc. CVPR. Volume 1. (2005) 984–989
9. Urtasun, R., Fua, P.: Human Motion Models for Characterization and Recognition. In: Automated Face and Gesture Recognition. (2004)
10. Brand, M., Kettnaker, V.: Discovery and Segmentation of Activities in Video. IEEE Trans. on PAMI **22** (2000) 844–851
11. Bobick, A.F., Davis, J.W.: The Recognition of Human Movement Using Temporal Templates. IEEE Trans. on PAMI **23** (2001) 257–267
12. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. In: Proc. ICCV. (2005)
13. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. on PAMI **19** (1997) 711–720
14. Marcel, S.: (Dynamic Hand Posture Database: http://www-prima.inrialpes.fr/FGnet/data/10-Gesture/dhp_marcel.tar.gz)
15. Yacoob, Y., Black, M.J.: Parameterized Modeling and Recognition of Activities. CVIU **73** (1999) 232–247
16. E. Ayyappa: Normal human locomotion, part 1: Basic concepts and terminology. Journal of Prosthetics and Orthotics **9** (1997) 10–17
17. Collins, R.T., Liu, Y., Leordeanu, M.: Online Selection of Discriminative Tracking Features. IEEE Trans. on PAMI **27** (2005) 1631–1643
18. Han, J., Bhanu, B.: Statistical Feature Fusion for Gait-based Human Recognition. In: Proc. CVPR. Volume 2. (2004) 842–847
19. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley and Sons, New York (2001)