Common spatial patterns for real-time classification of human actions

Ronald Poppe
Human Media Interaction Group, University of Twente, The Netherlands

ABSTRACT

We present a discriminative approach to human action recognition. At the heart of our approach is the use of common spatial patterns (CSP), a spatial filter technique that transforms temporal feature data by using differences in variance between two classes. Such a transformation focuses on differences between classes, rather than on modeling each class individually. As a result, to distinguish between two classes, we can use simple distance metrics in the low-dimensional transformed space. The most likely class is found by pairwise evaluation of all discriminant functions, which can be done in real-time. Our image representations are silhouette boundary gradients, spatially binned into cells. We achieve scores of approximately 96% on the Weizmann human action dataset, and show that reasonable results can be obtained when training on only a single subject. We further compare our results with a recent examplar-based approach. Future work is aimed at combining our approach with automatic human detection.

INTRODUCTION

Automatic recognition of human actions from video is an important step towards the goal of automatic understanding of human behavior. This understanding has many potential applications, including improved human-computer interaction, video surveillance and automatic annotation and retrieval of stored video footage. In general, these applications demand classification of human movement into several broad categories. Real-time and robust processing is often an important requirement, while there is still some control over the recording conditions. For example, human-computer interfaces require direct interaction. Another example is surveillance in the area of domotica, where elderly people are monitored to enable them to live independently for a longer period of time.

In the development of a human action recognition algorithm, one issue is the type of image representation that is used. At the one extreme, bag-of-word approaches (Batra et al., 2007, Niebles and Fei-Fei, 2007) have been used. At the other extreme, pose information is used (e.g. Ali et al. (2007)). In this chapter, we assume that the location of a human figure in the image is known. While this might seem unrealistic, related work by Thurau (2007) and Zhu et al. (2006) shows that this detection can be performed reliably and within reasonable time. Recent work on human detection by Wu and Nevatia (2007) and Lin et al. (2008) even deals with partial observations, but we do not consider these here. To encode the observation of the human figure, we use a grid-based silhouette descriptor, where each cell is a histogram of oriented boundary points. This representation resembles the concept of histograms of oriented gradients (HOG, Dalal and Triggs (2005)), as it models the spatial relations, yet is able to generalize about local variations.

For classification, we learn simple functions that can discriminate between two classes. Our main contribution is the application of common spatial patterns (CSP), a spatial filter technique that transforms temporal feature data by using differences in variance between two classes. After applying CSP, the first components of the transformed feature space contain high temporal variance for one class, and low variance for the other class. This effect is opposite for the last components. For an unseen sequence, we calculate the histogram over time, using only a fraction – the first and last components – of the transformed space. Each action is represented by the mean of the histograms of all corresponding training sequences, which is a very compact but somewhat naive representation. A simple classifier distinguishes between the two classes. All discriminant functions are evaluated pairwise to find the most likely action class. This introduces a significant amount of noise over class labels but works well for the given task. Note that CSP can be used with any image descriptor that is encoded as a vector of a fixed size, for example a histogram of codeword frequencies.

We obtained competitive results on the publicly available Weizmann action dataset introduced in Blank et al. (2005). One advantage of our method is that we require relatively few training samples. In fact, despite considerable variation in action performance between persons, we obtain reasonable results when training on data from a single subject. Also, we avoid retraining all functions when adding a new class, as the discriminative functions are learned pairwise, instead of jointly over all classes. Another advantage is that our approach is fast. Training of our classification scheme takes well under 1 second for all actions, with unoptimized Matlab code on a standard PC. After calculating the image descriptors, which can be done efficiently using the integral image (Zhu et al. 2006), classification can be performed in real-time as only a moderate number of simple functions have to be evaluated.

In the next section, we discuss related work on action recognition from monocular video. Common spatial patterns, and the construction of the CSP classifiers, are discussed subsequently. We evaluate our approach on the Weizmann dataset and perform additional experiments to gain more insight into the strengths and limitations of our approach. Finally, we summarize our approach and compare our results to those that have previously been reported in literature. An early version of this chapter appeared as Poppe and Poel (2008).

RELATED WORK ON HUMAN ACTION RECOGNITION

There has been a considerate amount of research into the recognition and understanding of human actions and activities from video and motion capture data. We discuss related work on action recognition, obtained from segmented monocular video sequences. More comprehensive overviews appear in Hu et al. (2004) and Turaga et al. (2008).

Action recognition can be thought of as the process of classifying arbitrary feature streams obtained from video sequences. This reveals the two main components of the problem: the feature representation and the classification. There have been many variations of both.

The choice of feature representation is important as it partly captures the variation in human pose, body dimension, appearance, clothing, and environmental factors such as lighting conditions. An ideal representation would be able to discriminate between poses, while at the same time being able to generalize over other factors. Since it is difficult to robustly obtain rich descriptors from video, often a compromise is sought in the complexity of the representation. At the one end, many approaches use retinotopic representations where the person is localized in the image. The image observations, such as silhouette or edge representations, are conveniently encoded into a feature vector. At the other end, there is the bag-of-words approach, where the

spatial dimension is ignored altogether. Feature representations that are somewhere in between these concepts, such as grid-based descriptors (e.g. Ikizler and Duygulu (2007), Wang and Suter (2007)), are currently popular. These encode the image observation locally as a bag-of-words, but preserve the spatial arrangement of these local descriptors.

Regarding classifiers, we can generally distinguish two large classes of action recognition approaches. Spatio-temporal templates match unseen sequences to known action templates. These templates can take many forms. A key frame or the mean of a sequence of silhouettes over time can be used as templates. Slightly more advanced is the concept of Motion History Images, introduced by Bobick and Davis (2001). Here, the differences between subsequent silhouettes are used, and stored in a two-dimensional histogram. Recent work by Blank et al. (2005) concatenates silhouettes over time to form a space-time shape. Special shape properties are extracted from the Poisson solution, and used for shape representation and classification.

The time dimension plays an important role in the recognition of actions, since there is often variation in the timing and speed with which an action is performed. Spatio-temporal templates can be considered as prototypes for a given action. The temporal aspect is often poorly modeled, especially when using histograms.

State-based representations, the second class of action classifiers, model the temporal aspect more accurately. These methods are often represented as a graphical model, where inference is used to perform the classification. Temporal relations between different states are encoded as transition probabilities. Hidden Markov Models (HMM) have been used initially (Brand, 1997). HMMs are also used by Weinland et al. (2007) for action recognition from arbitrary, monocular views. A similar approach using Action Nets was taken by Lv and Nevatia (2007).

Generative models try to maximize the likelihood of observing any example of a given class. For different actions that show many similarities yet have significant intra-class variance in performance (e.g. walking and jogging), generative models do a poor job in the classification task. Another drawback of generative models is the assumption that observations are independent. Discriminative models such as conditional random fields (CRF) condition on the observation, which makes this independence assumption unnecessary. Such models can model long-range dependencies between observations, as well as overlapping features.

Recently, discriminative alternatives have been proposed, based on CRFs. Sminchisescu et al. (2006) use CRFs and Maximum Entropy Markov Models (MEMM) to learn models for different actions simultaneously from image observations or motion capture data. Wang and Suter (2007) employ factorial conditional random fields (FCRF). Quattoni et al. (2007) use hidden conditional random fields (HCRF) that model the substructure of an action in hidden states. State-based approaches usually have a large number of parameters that need to be determined during training. This requires a large amount of training data, which is not always available.

In our approach, we learn functions that discriminate between two classes. Yet we avoid having to estimate a large number of parameters by representing actions as single prototypes. These prototypes lie in a space that is transformed by applying common spatial patterns on the feature data which are HOG-like representations of silhouette boundaries. We reduce the dimensionality of the feature representation and select the components that maximize the variance between the two classes. For an unseen action, we evaluate all pairwise discriminant functions, where each function softly votes into the two classes. Our estimated class label corresponds to the action that received most of the votes. Even though such an approach inherently generates a lot of noise in the classification, we show that we can recognize actions accurately, even when few training sequences are used. We explain common spatial patterns in the next section.

COMMON SPATIAL PATTERNS

Common Spatial Patterns (CSP) is a spatial filter technique that is often used in classifying brain signals (Müller-Gerking et al., 1999). It transforms temporal feature data by using differences in variance between two classes. After applying the CSP, the first components of the transformed data have high temporal variance for one class and low temporal variance for the other. For the last data components, this effect is opposite. When transforming the feature data of an unseen sequence, the temporal variance in the first and last components can be used to discriminate between the two classes. Consider the case where we have training sequences for two actions, *a* and *b*. Each training sequence can be seen as $n \times m_p$ matrix, where *n* is the number of features and m_p is number of time samples. We assume that the data is normalized in such a way that the mean of each feature is 0. Let C_a be the concatenation of the examples of action *a*, C_a is an $n \times m_a$ matrix. We do the same for action *b* to construct the matrix C_b. Now consider the matrix:

$$C = C_a C_a^T + C_b C_b^T \qquad (1)$$

C is the variance of the union of the two data sets. Since *C* is symmetric, there exists a orthogonal linear transformation *U* such that $? = UCU^T$, a positive diagonal matrix. The next step is to apply the whitening transformation $? = ?^{-1/2}$, which gives us $(? \ U)C(? \ U)^T = I$, and thus:

$$S_a \quad ? \quad (? \ U)C_a C_a^T (? \ U)^T \qquad (2)$$

$$S_b \quad ? \quad (? \ U)C_b C_b^T (? \ U)^T \qquad (3)$$

$$S_a \ ? \ S_b \quad ? \quad I \qquad (4)$$

Since S_a is symmetric, there is an orthogonal transformation *D* such that DS_aD^T is a diagonal matrix with decreasing eigenvalues on the diagonal. Hence, DS_bD^T is also a diagonal matrix but with increasing eigenvalues on the diagonal. The CSP is the spatial transform $W = D? U$ which transforms a data sequence into a sequence of dimension $2k$ such that a vector belonging to one action has high values in the first *k* components. For a vector of the other action, the situation is opposite. Hence, the temporal variance in these first and last components can be used to discriminate between action *a* and *b*.

CSP classifiers

Based on the CSP technique, we designed discriminating functions $g_{a,b}$ for every action *a* and *b* with $a \ ? \ b$. First we calculated the CSP transformation $W_{a,b}$ as described above. Then we applied $W_{a,b}$ to each action sequence of class *a* and *b*. Afterwards, the variance was taken over the entire sequence. This resulted in a single *n*-dimensional vector which can be considered a histogram, normalized for the length of the sequence. Next, we calculated the means *a'* and *b'* of these training vectors for action *a* and *b*, respectively. In order to compute $g_{a,b}(x)$ for an unseen action sequence *x*, we used the same procedure and first apply $W_{a,b}$ to *x*. We then calculate the variance over time over all components, which gives a vector *x'* of length *n*. Finally, $g_{a,b}(x)$ is defined as:

$$g_{a,b}(x) \ ? \ \frac{\left\| b' - x' \right\| \ ? \ \left\| a' ? x' \right\|}{\left\| b' - x' \right\| \ ? \ \left\| a' ? x' \right\|} \qquad (5)$$

Here, $\| x \|$ denotes the vector length, or norm, of x. Evaluation of a this function gives a continuous output in the [-1, 1] interval. Note that $g_{a,b} + g_{b,a} = 0$. With a rescaling and transform into the [0, 1] domain, we could interpret these outputs as probabilities. However, since we assume equal prior probabilities for each class, we use our voting scheme for clarity. Also, we could have used different discriminative functions. For example, we could have kept the individual training vectors, instead of the mean. This would allow to better model intra-class variance. In this case, one could use Mahalanobis distance, or use a margin classifier such as Support Vector Machine (SVM). These alternatives are, however, sensitive to outliers in the data.

We combined our pairwise classifiers into a multi-class classifier using voting. Such a scheme has been proposed by Friedman (1996) for binary outputs, i.e. $g'_{a,b}(x) = sgn(g_{a,b})$. We applied their work for continuous outputs, without loss of generality. In such a scheme, an action sequence is classified by evaluating all discriminant functions between pairs of a and b over all actions, and summing their votes:

$$g_a(x) \to \sum_{a \neq b} g_{a,b}(x) \qquad (6)$$

Since each action class appears in the exact same number of discriminative functions, the classification of x is the action $?$ for which $g_a(x)$ is maximal. This is the class that received most of the voting mass:

$$?(x) \to \arg\max_a g_a(x) \qquad (7)$$

Note that we also evaluate the discriminant functions in which the actual class does not appear. This introduces a large component of noise into the voting. However, actions that show more similarities with the unseen sequence will receive more mass in the voting. Hastie and Tibshirani (1998) remark that such a voting approach tends to favor classes that are closer to the average value in feature space. Such an effect would be larger for weaker pairwise discriminative functions. In our experiments, the dimensionality is relatively high compared to the number of classes and we expect that the effect of this bias is small.

More complex classification schemes are also possible. For example, Hastie and Tibshirani (1998) take into account all individual pairwise probability estimates and minimize a Kullback-Leibler criterion to find the optimal decision boundaries. The advantage is that the decision boundaries are determined jointly for all pairs of classes. Works by Allwein et al. (2000) and Dietterich and Bakiri (1995) use error-correcting codes, where each 'bit' in the code corresponds to a pairwise decision. While these approaches are better at dealing with noise caused by incidental erroneous decisions, their added value in performance over voting is limited (Allwein et al., 2000). Moreover, we prefer the straightforward interpretation of the voting outcome.

Our multi-class classifier requires $m(m - 1)/2$ functions, with m being the number of classes. Note that we could alternatively have used a one-versus-all classification scheme. In this case, we would have needed to learn a discriminative function for each class. While the complexity of such an approach is linear in the number of classes, instead of quadratic as in our scheme, the discriminative functions need to be more complex.

HISTOGRAMS OF ORIENTED SILHOUETTE GRADIENTS

To encode our image observations, we used a grid-based approach. For action recognition, grids were used as an image representation by Ikizler and Duygulu (2007), Thurau (2007) and Wang and Suter (2007). Our image representation is a variant of histogram of oriented gradients (HOG, Dalal and Triggs (2005)). For completeness, we summarize the processing steps used to obtain these descriptors.

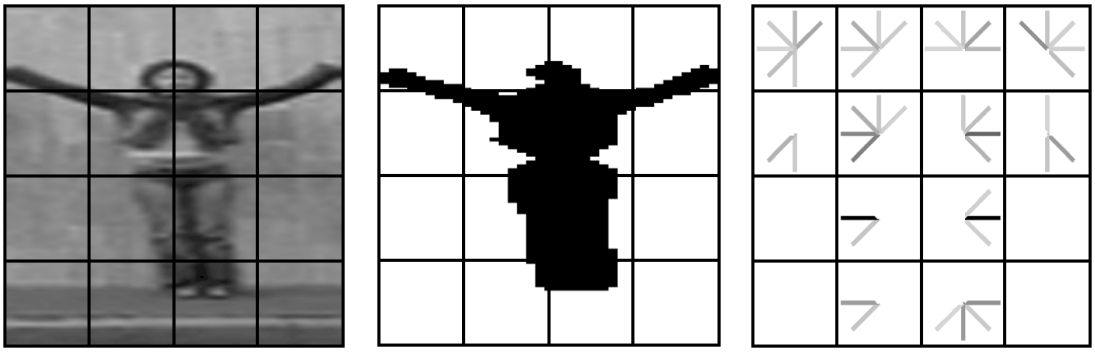

Figure 1: Silhouette descriptor, (left) image, (center) mask and (right) the boundary orientations, spatially binned into cells. Normal vectors are shown for clarity.

The different steps in our approach are shown in Figure 1. Given an extracted silhouette, we determine the enclosing bounding box, which determines the region of interest (ROI). We add space to make sure the height is 2.5 times the width. Next, we divide the ROI into a grid of 4×4 cells. Within each cell, we calculate the distribution of silhouette gradients, which we divide over 8 bins that each cover a 45° range. Pixels that are not on silhouette boundaries are ignored.

This idea is similar to that of histogram of oriented gradients (HOG) but our implementation is a simplification at a number of levels. First, we do not apply a Gaussian filter to enhance the edges. Second, we do not use overlapping cells, which significantly reduces the size of our descriptor. Third, and most important, we only take into account the silhouette outline thus discarding the internal edges. The fact that the gradient of binary silhouettes can only be vertical, horizontal or diagonal motivates the use of 45° orientation ranges. The final 128-dimensional descriptor is a concatenation of the histograms of all cells, normalized to unit length to accommodate variations in scale. We will refer to this representation as histograms of oriented silhouette gradients (HOSG). We will, however, also report the performance of our algorithm on different HOG and HOSG settings in our additional experiments.

Due to the normalization of the descriptor to unit length, and the relatively high dimensionality compared to the number of data points in a sequence, the covariance over a sequence may be nearly singular in some cases. We avoid this by applying PCA and select the 50 first components. These explain approximately 75% of the variance, depending on the subject that is left out. See the next section for details regarding this process.

EXPERIMENT RESULTS

We evaluated our approach on the publicly available Weizmann human action dataset which is briefly described in the next section. We then present the setup of our experiments and our obtained results. Additional experiments are described subsequently. A discussion of our results and a comparison with related work are given in the "Discussion" section.

Weizmann human action dataset

For the evaluation of our approach, we used the Weizmann action dataset (Blank et al., 2005, Gorelick et al., 2007). This set consists of 10 different actions, each performed by 9 different subjects (see also Figure 2). For subject Lena, additional sequences appear for the run, skip and walk action. We decided to leave these out in order to obtain a balanced set. This also allows for direct comparison of our results to those previously reported on the dataset. Note that our approach also works for unbalanced sets. The skip action was not originally present in the set and we present results both with and without the skip action.

Each sequence is approximately 2.5 seconds long. There is considerable intra-class variation due to different performances of the same action by different subjects. Most notably, the jump, run, side, skip and walk actions are performed either from left to right, or in the opposite direction. Since the actions were performed on a slight slope, the direction of movement also results in slightly different movement style. Despite these differences, we treated performances in different directions as belonging to the same class. The sequences were recorded from a single camera view, against a static background, with minimal lighting differences. Binary silhouette masks are provided with the dataset. These masks contain a considerable amount of noise due to inaccurate background segmentation (see also Figure 6).

Figure 2: Example frames from the Weizmann dataset. Different subjects performing actions bend, jack, jump, pjump, run, side, skip, walk, wave1 and wave2.

Experiment setup

We evaluated our method using leave-one-out cross-validation (LOOCV), where each of the 9 folds corresponds to all sequences of the corresponding subject. Specifically, this gave us 80 training sequences per fold, 8 for each of the 10 actions. First, we calculated the PCA transformation over all training sequences and projected the silhouette descriptors onto the first 50 components. Next, we learned all discriminant functions $g_{a,b}$, between all pairs of actions a and b ($1 = a$, $b = 10$, $a ? b$). Specifically, we used the first and last $k = 5$ components in the transformation, which gave us action prototypes vectors of dimension 10. We experimented with other values for k but found no improvement for $k > 5$. For each of the sequences of the subject whose sequences were left out, we evaluated all discriminant functions. Each of these evaluations softly votes over class a and b. In our final classification, we selected the class that received the highest voting mass.

Results

We performed the LOOCV experiment and obtained a performance of 95.56%. In total, 4 sequences were misclassified. The skip action of subject Daria was classified as jumping, the skip action of subject Ido was classified as running. Also, the jump action of subject Eli and the run action of subject Shahar were both classified as walking. The confusion matrix for this experiment is shown in Table 1.

In order to be able to compare our results with those reported in previous studies, we also left out the skip class. This resulted in a performance of 96.30%. Again, the jump action of subject Eli and the run action of subject Shahar were classified as walking. In addition, the wave1 action of subject Lyova was misclassified as wave2.

In line with Friedman (1996), we also evaluated the performance when using binary outputs for the discriminative functions (i.e. $g'_{a,b}(x) = sgn(g_{a,b})$). With the skip action, 3 additional errors were made which resulted in a performance of 92.22%. Without skip, the performance was similar to the soft vote case at 96.30%.

	Guessed									
Actual	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	9									
Jack		9								
Jump			8					1		
Pjump				9						
Run					8			1		
Side						9				
Skip			1		1		7			
Walk								9		
Wave1									9	
Wave2										9

Table 1: Confusion matrix for Weizmann dataset including skip action with CSP (performance 95.56%). See text for explanation.

	Guessed									
Actual	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	7	1							1	
Jack		9								
Jump		1	5			1	1		1	
Pjump				9						
Run					6			2	1	
Side						9				
Skip		1			3		2	3		
Walk					1		1	7		
Wave1									8	1
Wave2		1								8

Table 2: Confusion matrix for Weizmann dataset including skip action without CSP (performance 77.78%). See text for explanation.

Both the feature representation and the classifier had an important impact on the performance. To measure the added value of using CSP, we performed an additional experiment where we did not transform the feature space. Instead, we took the first 10 components of the PCA. For each training sequence, we calculated the histogram by taking the mean of the feature vector over time which resulted in a 10-dimensional vector. We determined the prototype for each action by averaging these histograms. Again, we used Equations (5) and (7) to determine the class estimate. We achieved a performance of 77.78% for all actions, and 85.19% with the skip action omitted. The confusion matrix for all 10 actions is shown in Table 2. When we used the first 50 PCA components, the performance slightly increased to 80.00% for all actions, while the performance without skip remained the same. A closer look at the misclassifications shows confusion between run, skip and walk, along with some incidental confusions. It thus becomes clear that the use of CSP is advantageous over a feature representation without CSP transform.

The baseline for the full dataset is 10.00%, and 11.11% when the skip action is left out. Obviously, our results are well above these baselines and show that we can achieve good recognition, even when single action prototypes of dimension 10 are used. Also, it shows that intra-class variations can be handled without modeling the variance between different subjects. To gain insight in the characteristics of our method, we conducted additional experiments. These are described in the next section.

ADDITIONAL EXPERIMENTS

In addition to the evaluations described above, we conducted several additional experiments to see how our approach performs with different settings and under different conditions. We used our HOSG descriptors with the settings as described previously, unless stated otherwise. Also, we used the standard Weizmann dataset, except for the robustness experiment.

First, we present our experiment with different image representations. Next, we describe our experiments where we used only part of the available training data. Evaluations on sequences with different deformations and viewpoints are then discussed. Finally, we describe our experiment with recognition from a smaller number of frames.

Results using different image representations

In this section, we evaluate the effect of descriptor size and type on the classification performance. Specifically, we used 2 descriptor types: HOG and HOSG. The former uses edges extracted within a silhouette mask. We also used three different grid sizes: 3 × 3, 4 × 4 and 5 × 6. HOSG-4 × 4 was used in the previous section. Descriptor sizes for HOG are 81, 144 and 270 for the grid sizes respectively. For HOSG, these sizes are 72, 128 and 240. We kept the number of CSP components constant. Unseen sequences and action prototypes were both points in 10-dimensional space ($k = 5$).

We used the LOOCV approach for evaluation, with the data of 8 subjects for training and the data of the remaining subject for testing. The results are summarized in Table 3. HOSG performed slightly better than HOG. We can see that 4 × 4 outperformed both smaller and bigger grids. We expect that 3 × 3 grids do not capture enough detail to distinguish between classes. For 5 × 6 grids, we contribute the lower performance to the smaller cell sizes. This causes the histograms to become sparse which results in higher similarity scores when small variations between performances of an action occur.

	HOSG			HOG		
	3 × 3	4 × 4	5 × 5	3 × 3	4 × 4	5 × 5
All actions	84.44%	95.56%	85.56%	83.33%	90.00%	87.78%
Skip omitted	88.89%	96.30%	91.36%	90.12%	92.59%	90.12%

Table 3: Classification performance using HOSG and HOG for different grid sizes.

Results using less training data

The fact that our approach is able to perform well, even though intra-class variation is not modeled, gives the impression that we can train our classifiers with less training data. Note here that training for all actions and all training subjects takes well under 1 second. To verify this hypothesis, we evaluated the performance of our approach using different numbers of subjects in the training set. Again, we used the LOOCV scheme. For each number of training subjects *k*, we present the results as averages over all *8!/(k!(8-k)!)* combinations of training subjects. Table 4 summarizes these results, both using all actions, and with the skip action omitted.

Number of subjects	Number of combinations	All actions	Skip omitted
1	8	64.72%	69.14%
2	28	77.82%	83.82%
3	56	81.83%	88.98%
4	70	84.60%	90.85%
5	56	86.63%	92.44%
6	28	89.01%	93.87%
7	8	91.39%	94.91%
8	1	95.56%	96.30%

Table 4: Classification performance of our CSP classifier on the Weizmann dataset, using different numbers of training subjects. Combinations is the evaluated number of subsets of subjects.

Clearly, performance decreases with a decreasing amount of training data. But, even when only a few subjects are used for training, the results are reasonable. We expect that the variation in the direction of movement of the jump, run, side, skip and walk sequences will have a significant impact on the results, especially for the evaluations with very few training subjects. Even though we do not model the movement in the image, changing the direction of movement results in mirrored image observations. In turn, this results in very different silhouette descriptors. We look at this issue further in the next section. Nevertheless, our approach can cope with these variations to some extent.

Results on robustness sequences

The Weizmann dataset contains additional robustness sequences that can be used to investigate how well an approach performs with more challenging data. There are two types of sets, each of which contain 10 additional walking sequences. In the deformation sequences, different variations of walking are viewed from the side (see Figure 3 (top row)). These sequences include walking with objects (bag, briefcase, dog), different walking styles (kneesup, limp, moonwalk), different clothing styles (skirt) and occlusion settings (nofeet, pole). It is arguable whether the different styles should be classified as walking since they show many similarities with the skip

action. The viewpoint sequences show one walking subject, viewed from 0° (side view) to 81° (near-front view), in increments of 9°. Figure 1 (bottom row) shows example frames.

Our experimental setup was similar to the one used earlier but we used training data of all 9 subjects. We performed the experiments on the deformation and viewpoint sequences separately. We used the HOSG-4 × 4 descriptors. Our results are averaged over the 10 sequences of each set. For the deformation sequences, we obtained 80.00% correct estimates. The incorrectly classified sequences were moonwalk and pole, both of which were classified as running. For the viewpoints sequences, 80.00% were also classified correctly. The most challenging trials corresponding to viewpoints 72° and 81° were both classified as pjump.

Figure 3: Example frames from the Weizmann robustness sequences. (top) Deformations, images and silhouettes for bag, briefcase, dog, kneesup, limp, moonwalk, nofeet, normwalk, pole and skirt. (bottom) Different viewpoints, images and silhouettes, 0° - 81° in increments of 9°.

When we reduced the number of subjects in our training set, we obtained lower results. Specifically, for 5 subjects, we score 79.60% correctly on the deformations, and 70.16% on the viewpoints. For training on a single subject, these numbers decrease to 58.89% and 48.89%, respectively. These percentages are averages of all combinations of training subjects. For the condition where we test only on a single subject, we can evaluate the influence on walking direction on the performance, as the sequences in both the deformations and viewpoints sets show walking from left to right. When the training subject is walking in the same direction as the test subject, the scores are respectively 80.00% and 74.00% on the deformations and viewpoints sets. For the opposite direction, these numbers are significantly lower at 32.50% and 17.50%, respectively. Here, we did not look at the direction of related classes such as run and skip but it shows that it is important to take the direction of movement into account during training. Alternatively, different directions can be treated as different action classes.

Results on subsequences

So far, we have used the entire sequence for classification. We assumed that temporal segmentation was performed previously. This raises the question as to how well our approach would perform when such accurate segmentation is not available. Since the Weizmann dataset contains only sequences with a single action, we focus on subsequences instead. We repeated our main LOOCV experiment but varied the length of the test sequences. The training phase was exactly the same, so we used the entire sequences. For testing, we used a sliding window with a length in the range [1, 25]. The minimum sequence length is 28 frames. We slid the window through the sequence with steps of 1 frame. Average performance results over all sequences for

different subsequence lengths are given in Figure 4 (left). It is clear that increasing subsequence length results in an increased performance. This can be explained by the additional information that is available as the sequence becomes longer.

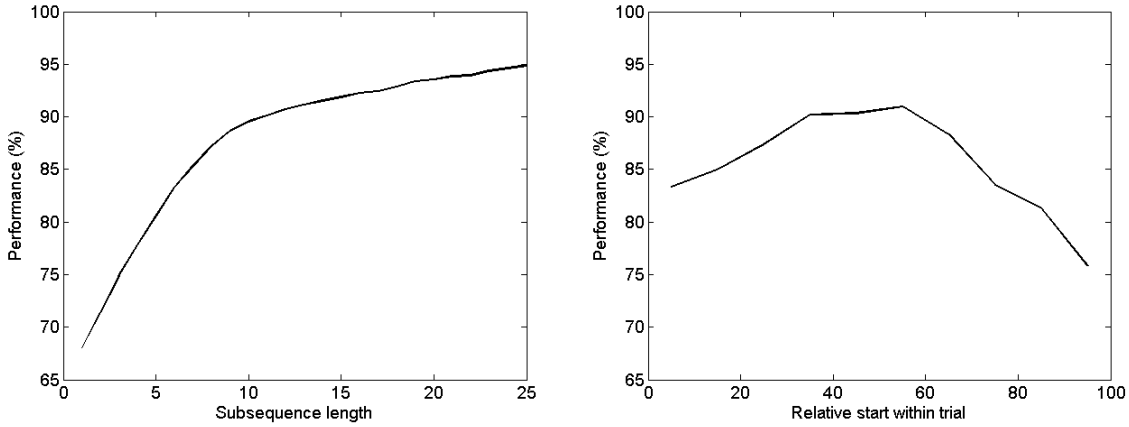

Figure 4: Classification performance for different subsequence lengths (left) and at different relative times (percentages given) within the sequence (right).

We expect that the relative progress within the sequence influences these results. Most action performances start and end in a resting pose. Also, for moving actions (e.g. walking and running), the start and end of the sequence take place partly outside the viewing window. Therefore, we looked at the classification performance at different relative times within the sequence. We calculated the relative start of the subsequence as the starting frame divided by the sequence length. To compare sequences of different lengths, we binned these values into a 10-dimensional histogram. Each cell contains the average classification performance. Figure 4 (right) shows these results, averaged over all sequence lengths. It immediately becomes clear that performance is indeed lower at the start and at the end of the sequence.

DISCUSSION

In this section, we compare our results with those reported previously in literature. In the first sub-section, we present an in-depth comparison with recent exemplar-based holistic work. Next, we compare our approach with other results on the Weizmann human action dataset. Finally, we summarize our approach and present directions for future research.

Comparison with exemplar-based holistic work

In many cases, humans can recognize human actions from only a single prototypical pose. Motivated by this observation, we explored the use of such key poses. Recently, Weinland and Boyer (2008) presented an approach where they described sequences as a vector of minimum distances to selected exemplars. There are several approaches to select these exemplars. Unsupervised clustering algorithms such as k-means and expectation-maximization (Dempster et al., 1977) are likely to select as exemplars those frames that are common among all classes. As such, they are not discriminative. Alternatively, the exemplar selection problem can be regarded as a feature subset selection problem, where each frame is a feature. There are three types of supervised approach (Blum and Langley, 1997, Guyon and Elisseeff, 2003). Filters select subsets

as a preprocessing step, without taking into account the induction algorithm (classifier). Wilson and Martinez (2000) present an overview of filter approaches. In contrast, wrapper approaches (Kohavi and John, 1997) explicitly use the induction algorithm in the subset selection scheme. A third approach is that of embedding methods, which perform subset selection within the training process. Usually, these methods are specifically designed for a given classifier and we do not consider them here.

In this section, we describe our implementation of the approach of Weinland and Boyer (2008), using either k-medoids (k-means where cluster centra correspond to the closest exemplar) or the wrapper approach to select exemplars. We used a Bayes classifier where each class is described as a multivariate Gaussian. Given the conceptual advantages of the wrapper approach over the unsupervised k-medoids, we expect to achieve higher accuracies for a smaller number of exemplars.

We used a LOOCV approach where each fold corresponds to one of the 9 test subjects. Our settings corresponded to those in Weinland and Boyer, (2008), which we summarize here for completeness. Specifically, we used a forward selection scheme. We started with an empty set of exemplars $E = \emptyset$, and a full set of candidates $C = \{c_i \mid 1 = i = n\}$ with n the total number of candidates. We sampled $n = 300$ frames from the training sequences. At each iteration, an exemplar from the candidate set was moved to the exemplar set. This was the exemplar that resulted in the largest performance increase on the validation set. To make sure exemplar selection was not biased on a single subject, we used cross-validation within this exemplar-selection step. Since a perfect performance score on the validation set is easily obtained, we temporarily and randomly removed exemplars until the validation score was below 100%. In the validation step, we used the Bayes classifier where each class was described as a multivariate Gaussian. To avoid singularity problems in the inversion, we used an axis-aligned covariance matrix, in which all off-diagonal covariance elements are zero. We used Mahalanobis distance D to determine the distance of each unseen trial to all classes: $D = (x - \mu)^T S^{-1}(x - \mu)$, where x is the k-dimensional vector of minimum distances to the k selected exemplars, and μ and S are the mean and covariance of the given class, respectively.

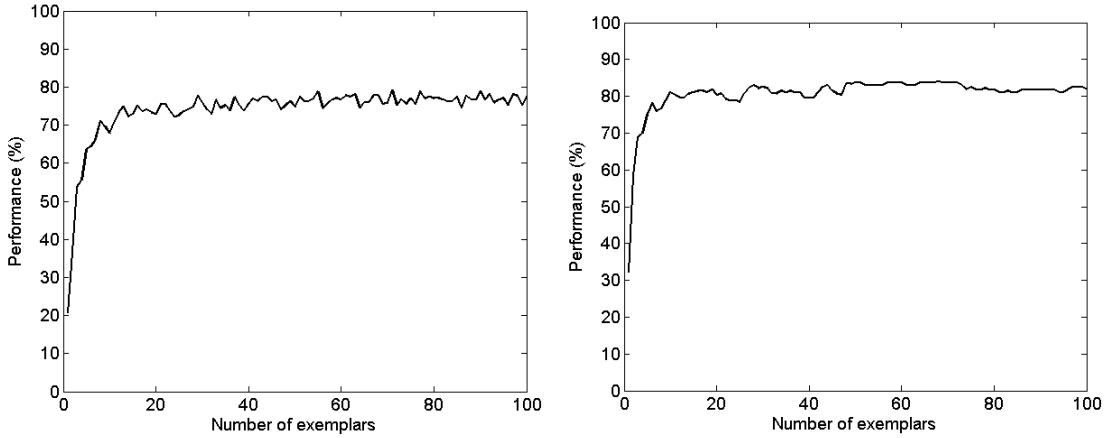

Figure 5: Classification performance for different numbers of exemplars k *for k-medoids (left) and the wrapper approach (right).*

When multiple frames resulted in the highest performance increase on the validation set, we randomly selected one of them. Also, the selection of the candidate set was random. Therefore,

we present our results as averages over 3 repetitions. Again, we used the HOSG-4 × 4 descriptor as our image representation, and performed our experiments with all 10 action classes.

The results for different numbers of *k* are presented in Figure 5, with either k-medoids or the wrapper approach for exemplar selection. The graphs show that for the wrapper approach, performance increases more rapidly. This can be understood by the discriminative selection in the wrapper approach. Also, the performance with the wrapper approach is slightly higher. For one repetition, the exemplars for *k = 10* are shown in Figure 6. The exemplars that are selected in the wrapper approach correspond more clearly to different classes, whereas k-medoids selects exemplars that are more common among classes. Confusion matrices for *k = 50* are presented in Table 5. It is clear that run and skip are often guessed, which can be explained by the large within-variance. Remarkably, the two wave actions are both often classified as bend. This is probably due to scaling the bounding box to a fixed ratio. Notice the perfect recognition for the walk action when the wrapper approach is used. This shows the discriminative effect of the selected exemplars (see exemplar 1 and 7 in the bottom row of Figure 6).

Figure 6: Exemplars selected using k-medoids (top), and the wrapper approach (bottom), both with the number of exemplars k = 10. *Test subject is Daria.*

Both the exemplar-based approach and our CSP classifiers are discriminative but their strengths are different. The results of the exemplar-based approach are easily interpretable and arbitrary distance measures between frames and exemplars can be used. For example, Weinland and Boyer (2008) use Chamfer distance and achieve 100% accuracy when at least 120 exemplars are used. The CSP classifiers are limited in that they require a vector representation. However, the CSP classifiers can be trained very efficiently and have been shown to yield good results even for small subsets or when limited training data is available. In a direct comparison using the HOSG descriptors our CSP classifier outperforms the exemplar-based approach with over 10%. Differences between our results on the wrapper approach and those reported in Weinland and Boyer (2008) can be explained by the different image representation and matching. The Chamfer matching is more robust at the cost of being more computationally expensive.

Comparison with other related research

There have been several other reports of results on the Weizmann set. We review these and point out differences with our work. Such comparisons reveal the relative advantages of one method over the other. We selected works that are representative of a class of approaches.

Niebles and Fei-Fei (2007) achieved a 72.80% score over 9 actions. Spatial and spatiotemporal interest points were sampled, and combined into a constellation. Action classification was performed by taking a majority vote over all individually classified frames. No background segmentation or localization was needed. This makes their approach more robust

than ours. Recent work by Thurau (2007) used HOG-descriptors for both detection and action classification. No background segmentation was used, but centered and aligned training data was needed. For classification, n-grams of action snippets were used. With all 10 actions and 90 bi-grams, performance was 86.66%.

	Guessed									
Actual	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	23				3		1			
Jack		25			2					
Jump	2		15		6		4			
Pjump				24	3					
Run					24		3			
Side					3	24				
Skip					18		9			
Walk					10		1	16		
Wave1	6				1				20	
Wave2	4				1					22

	Guessed									
Actual	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	24						3			
Jack		27								
Jump			18				9			
Pjump				25	2					
Run					24		3			
Side					3	24				
Skip					14		13			
Walk								27		
Wave1	6								21	
Wave2	3					1				23

Table 5: Confusion matrix for exemplar-based experiment with HOSG descriptors, with k = 50. Exemplars are selected using k-medoids (top, performance 74.81%) or the wrapper approach (bottom, performance 83.70%). The numbers are accumulated for 3 repetitions.

In theory, the work of Ikizler and Duygulu (2007) did not require background segmentation but localization was assumed. A large number of rotated rectangular patches were extracted, and divided over a 3 × 3 grid, forming a histogram of oriented rectangles. A number of settings and classification methods was evaluated on the dataset without the skip action. All actions were classified correctly when using Dynamic Time Warping. This requires the temporal alignment of each unseen sequence to all sequences in the training set, which is computationally expensive. Using one histogram per sequence, 96.30% was scored. Again, this required comparison to all training sequences. For comparison, we calculated the performance of our descriptor using a length-normalized histogram over the entire sequence and 1-nearest neighbor using Euclidian distance, and with the skip action left out. This resulted in 96.30% performance, a similar score.

Other works require background subtraction and use the masks that are provided with the dataset. Wang and Suter (2007) score 97.78% over all 10 actions. Raw silhouette values were used, and long-term dependencies between observations were modeled in their FCRF. When small blocks of pixels were regarded, thus effectively reducing the resolution, performance decreased. For 4 × 4 blocks and 8 × 8 blocks, scores were obtained of 92.22% and 77.78%, with descriptor sizes 192 and 48, respectively. Kernel PCA was used to reduce the dimensionality, but the dimension of the projected space was not reported. In contrast, we started with a 128-dimensional silhouette descriptor, and performed the classification using only 10 components. Moreover, our training requirements were much lower. On the other hand, FCRFs are able to model complex temporal dynamics.

There are several reports of subsequence classifications. For example, Blank et al. (2005) used subsequences of 10 frames, and obtained a performance of 99.64%. They used local features, extracted from a space-time volume that was constructed by concatenating silhouettes over time. Schindler and Van Gool (2007) used local shape and optical flow, and evaluated their approach using subsequences between one and 10 frames. Their performance of 93.5% for a single frame increased to 99.60% when 10 frames were used. In contrast to these works, we used a holistic representation and no motion information. Such a representation can be obtained much faster. The downside is our lower performance of 89.56% using 10-frame subsequences.

Conclusion

We have shown that the application of common spatial patterns (CSP) to increase the margin between pairs of classes, increases classification performance. We demonstrated our approach on the Weizmann dataset and obtained approximately 96% accuracy. Confusions that remain are between related classes such as walking and running. These results are competitive, and we have shown that we can obtain reasonable results with only a few training subjects. Moreover, training and evaluation complexity are low. In fact, we can perform human action recognition in real-time. CSP can take sequences of any fixed-size vector representation as input. Here, we have used histograms of oriented silhouette gradients (HOSG), calculated within a grid. Such a holistic representation can be calculated fast but requires background segmentation and the determination of the region of interest. To assess the performance of our method on more realistic scenes reliably, our work should be combined with a preprocessing step to automatic human detection, such as in Zhu et al. (2006). In situations where silhouettes cannot be obtained reliably, histograms of codeword frequencies, for example from interest point detectors, can be used.

For the classification, we used simple pairwise discriminative functions, where each class was represented by the average of all training sequences of the class. Such an approach is simple, but does not model intra-class variance. Each prototype is likely to be an average of multiple modes, especially when there are large differences within the class, such as different directions of movement. To overcome this issue, multiple classes for different direction of movement for a single action could be introduced. Moreover, the temporal aspect in our action prototypes is, to a large extent, ignored. Performance could be increased by including temporal characteristics.

We have evaluated our work on entire sequences and subsequences. We did not explicitly address the temporal segmentation. Also, current datasets for human action recognition do not contain an “other” class. Instead of selecting the class with the highest voting mass, this would also require an approach to decide whether the chosen class is really observed. Generally, this is a harder problem since there is more variation in the “other” class and the prior probabilities for the classes can vary significantly.

ACKNOWLEDGEMENTS

The author would like to thank Mannes Poel and Daniel Weinland for insightful discussions on parts of this chapter, and the authors of Blank et al. (2005) for making their dataset available. This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024.

REFERENCES

Ali, S., Basharat, A., and Shah, M. (2007). Chaotic invariants for human action recognition. In Proceedings of the International Conference On Computer Vision (ICCV'07), pages 1–8, Rio de Janeiro, Brazil.

Allwein, E. L., Schapire, R. E., and Singer, Y. (2000). Reducing multiclass to binary: a unifying approach for margin classifiers. The Journal of Machine Learning Research, 1:113–141.

Batra, D., Chen, T., and Sukthankar, R. (2008). Space-time shapelets for action recognition. In Proceedings of the Workshop on Motion and Video Computing (WMVC'08), pages 1–6, Copper Mountain, CO.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 2, pages 1395–1402, Beijing, China.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97(1):245–271.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 23(3):257–267.

Brand, M., Oliver, N., and Pentland, A. P. (1997). Coupled hidden Markov models for complex action recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'97), pages 994–999, San Juan, Puerto Rico.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 1, pages 886–893, San Diego, CA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39(1):1–38.

Dietterich, T. G. and Bakiri, G. (1995). Solving multiclass learning problems via errorcorrecting output codes. Journal of Artificial Intelligence Research, 2:263–286.

Friedman, J. H. (1996). Another approach to polychotomous classification. Statistics department, Stanford University, Stanford, CA.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 29(12):2247–2253.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research (JMLR), 3:1157–1182.

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. Annals of Statistics, 26(2):451–471.

Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. IEEE Transactions On Systems, Man, And Cybernetics (SMC) - Part C: Applications And Reviews, 34(3):334–352.

Ikizler, N. and Duygulu, P. (2007). Human action recognition using distribution of oriented rectangular patches. In Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07), number 4814 in Lecture Notes in Computer Science, pages 271–284, Rio de Janeiro, Brazil.

Kohavi, R. and John, G. H. (1997). Wrappers for feature selection. Artificial Intelligence, 97(1):273–324.

Lin, Z. and Davis, L. S. (2008). A pose-invariant descriptor for human detection and segmentation. In Proceedings of the European Conference on Computer Vision (ECCV'08) - part 4, number 5305 in Lecture Notes in Computer Science, pages 423–436, Marseille, France.

Lv, F. and Nevatia, R. (2007). Single view human action recognition using key pose matching and Viterbi path searching. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), pages 1–8, Minneapolis, MN.

Müller-Gerking, J., Pfurtscheller, G., and Flyvbjerg, H. (1999). Designing optimal spatial filters for single-trial EEG classification in a movement task. Clinical Neurophysiology, 110(5):787–798.

Niebles, J. C. and Fei-Fei, L. (2007). A hierarchical model of shape and appearance for human action classification. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), pages 1–8, Minneapolis, MN.

Poppe, R. and Poel, M. (2008). Discriminative human action recognition using pairwise CSP classifiers. In Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08), Amsterdam, The Netherlands.

Quattoni, A., Wang, S. B., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 29(10):1848–1852.

Schindler, K. and Gool, L. J. van (2008). Action snippets: How many frames does human action recognition require? In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), pages 1–8, Anchorage, AK.

Sminchisescu, C., Kanaujia, A., and Metaxas, D. N. (2006). Conditional models for contextual human motion recognition. Computer Vision and Image Understanding (CVIU), 104(2-3):210–220.

Thurau, C. (2007). Behavior histograms for action recognition and human detection. In Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07), number 4814 in Lecture Notes in Computer Science, pages 271–284, Rio de Janeiro, Brazil.

Turaga, P., Chellappa, R., Subrahmanian, V., and Udrea, O. (2008). Machine recognition of human activities: A survey. IEEE Transactions On Circuits And Systems For Video Technology, 18(11):1473–1488.

Wang, L. and Suter, D. (2007). Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), pages 1–8, Minneapolis, MN.

Weinland, D., Boyer, E., and Ronfard, R. (2007). Action recognition from arbitrary views using 3D exemplars. In Proceedings of the International Conference On Computer Vision (ICCV'07), pages 1–8, Rio de Janeiro, Brazil.

Weinland, D. and Boyer, E. (2008). Action recognition using exemplar-based embedding. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), pages 1–7, Anchorage, AK.

Wilson, D. and Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. Machine Learning, 38(3):257–286.

Wu, B. and Nevatia, R. (2007). Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision (IJCV), 75(2):247–266.

Zhu, Q., Avidan, S., Yeh, M.-C., and Cheng, K.-T. (2006). Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2, pages 1491–1498, New York, NY.