Kernel-based Recognition of Human Actions Using Spatiotemporal Salient Points

A.Oikonomopoulos
Computing Department
Imperial College London
aoikonom@doc.ic.ac.uk

I.Patras
Computer Vision and
Pattern Recognition Group
The University of York
I.Patras@cs.york.ac.uk

M.Pantic
Computing Department
Imperial College London
m.pantic@imperial.ac.uk

Abstract

This paper addresses the problem of human action recognition by introducing a sparse representation of image sequences as a collection of spatiotemporal events that are localized at points that are salient both in space and time. We detect the spatiotemporal salient points by measuring the variations in the information content of pixel neighborhoods not only in space but also in time. We derive a suitable distance measure between the representations, which is based on the Chamfer distance, and we optimize this measure with respect to a number of temporal and scaling parameters. In this way we achieve invariance against scaling, while at the same time, we eliminate the temporal differences between the representations. We use Relevance Vector Machines (RVM) in order to address the classification problem. We propose new kernels for use by the RVM, which are specifically tailored to the proposed spatiotemporal salient point representation. The basis of these kernels is the optimized Chamfer distance of the previous step. We present results on real image sequences from a small database depicting people performing 19 aerobic exercises.

1. Introduction

Analysis and interpretation of image sequences plays an important role in the development of a wide range of vision-related applications, ranging from vision-based interfaces for Human-Computer Interaction (HCI), to surveillance and video indexing systems. Recognition and interpretation of human activities is a significant research area by itself, since a large amount of the information content of image sequences is carried in the human actions that are depicted in them. In order to obtain a semantic description of the content of a scene, we do not need to use all the available information. A good description of the scene can be obtained by considering the information around certain points of interest such as corners and edges, that is, in areas that are rich in information. According to Haralick and Shapiro [5] an interesting point is a) distinguishable from its neighbors and b) its position is invariant with respect to the expected geometric transformation and to radiometric distortions. Schmid et al. [16] detect interesting points using a Harris corner detector and estimate gray value differential image invariants [11],[19] at different scales. Gilles introduces the notion of saliency in terms of local signal complexity or unpredictability in [4]. Finally, detectors of interesting points are compared in [17],[18] in terms of repeatability rate and information content.

An important issue in salient point detection is the automatic selection of the scale at which the salient points will be detected. Lindeberg et al. [13] integrate a scale-space approach for corner detection and search for local extremes across scales. Itti et al [7] use a dyadic Gaussian pyramid approach in order to construct saliency maps from given images. The spatial distribution of each saliency map is modeled with a dynamical neural network, in order to select locations of attention in the images. Kadir and Brady [9],[8] extend the original Gilles algorithm and estimate the information content of pixels in circular neighborhoods at different scales in terms of the entropy. Local extremes of changes in the entropy across scales are detected and the saliency of each point at a certain scale is defined in terms of the entropy and its rate of change at the scale in question. In [6], the salient point detector developed in [9],[8] is compared with methods that use global descriptors in order to represent a given image, giving a clear advantage to the former method. Finally, an object recognition approach using keypoints is described by Lowe in [14]. The detected keypoints are invariant to geometric and illumination changes in the image. The spatial arrangement of the detected points is then used for retrieving objects from cluttered images.

While a large amount of work has been done on object

0-7695-2646-2/06 $20.00 (c) 2006 IEEE

recognition and image retrieval, the concept of saliency has only recently begun to be used for content-based video retrieval and for activity recognition. In [12], a Harris corner detector is extended in the temporal dimension, leading to a number of corner points in time, called space-time interest points. The resulting interesting points correspond roughly to points in space-time where the motion abruptly changes direction. The resulting representations are compared using a Mahalanobis distance metric. In [3], hand gestures are recognized by using a hierarchical hand model consisting of the palm and the fingers. Color features under different scales are then extracted and a particle filtering approach is implemented to simultaneously track and detect the different motion states occurring in the image sequence. In [10], visual interest points are computed from dynamically changing regions in a given image sequence. A selective control method is then used in order to equip the recognition system. In [1], given image sequences are used to construct Motion Energy Images (MEI) and Motion History Images (MHI), for determining where and when respectively motion occurs in the sequence. For recognition, a set of moment invariants is calculated for each resulting image and a Mahalanobis distance metric is applied in order to discriminate between different kinds of motion. In [21], a set of MEIs and MHIs is constructed from given image sequences for the task of facial action recognition. A k-Nearest-Neighbor Rule-based classifier is then utilized to detect the various Action Units (AUs) that are being activated in each facial expression.

In this paper, we detect spatiotemporal features in given image sequences by extending in the temporal direction the salient feature detector developed in [9],[8]. Our goal is to obtain a sparse representation of a human action as a set of spatiotemporal points that correspond to activity variation peaks. In contrast to the work of Laptev [12], our representation contains the spatiotemporal points at which there are peaks in activity variation such as the edges of a moving object. While in our previous work [15] we did so by examining the entropy in spherical spatiotemporal volumes, here we examine the entropy behaviour in spatiotemporal cylinders. By this, we aim in improving the final recognition rate that was reported in [15]. Like in [9], we automatically detect the scales at which the entropy achieves local maxima and form spatiotemporal salient regions by clustering spatiotemporal points with similar location and scale. Each image sequence is represented as a set of spatiotemporal salient points, the locations of which are transformed in order to achieve invariance against translation. We use Relevance Vector Machines in order to address the classification problem. We propose new kernels for use by the RVM, which are specifically tailored to the proposed spatiotemporal salient point representation. More specifically, we derive a suitable distance measure between the representations, based on the Chamfer distance, and we optimize this measure with respect to a number of temporal and scaling parameters. In this way we achieve invariance against scaling and we eliminate the temporal differences between the representations. This optimized distance is used as the basis for the definition of the kernel for the RVM. We test the proposed method using real image sequences. Our experimental results show fairly good discrimination between specific motion classes.

The remainder of the paper is organized as follows: In section 2, the spatiotemporal feature detector used is described in detail. In section 3 the proposed recognition method is analyzed, including the proposed space-time warping technique. In section 4, we present our experimental results, and in section 5, final conclusions are drawn.

2. Spatiotemporal Salient Points

2.1. Spatiotemporal Saliency

Let us denote by $N_c(s, \vec{v})$ the set of pixels in an image I that belong to a circular neighborhood of radius s, centered at pixel $\vec{v} = (x, y)$. In [9],[8], in order to detect salient points in static images, Kadir and Brady define a saliency metric $y_D(s, \vec{v})$ based on measuring changes in the information content of N_c for a set of different circular radiuses (i.e. scales). In order to detect spatiotemporal salient points at peaks of activity variation we extend the Kadir's detector by considering cylindrical spatiotemporal neighborhoods at different spatial radiuses s and temporal depths d, in a similar way as [15], where only spherical neighborhoods of radius s were used. More specifically, let us denote by $N_{cl}(\vec{s}, \vec{v})$ the set of pixels in a cylindrical neighborhood of scale $\vec{s} = (s, d)$ centered at the spatiotemporal point $\vec{v} = (x, y, t)$ in the given image sequence. At each point $\vec{v}$ and for each scale $\vec{s}$ we will define the spatiotemporal saliency $y_D(\vec{s}, \vec{v})$ by measuring the changes in the information content within $N_{cl}(\vec{s}, \vec{v})$. Since we are interested in activity within an image sequence, we consider as input signal the convolution of the intensity information with a first-order Gaussian derivative filter. Formally, given an image sequence $I_0(x, y, t)$ and a filter G_t, the input signal that we use is defined as:

$$I(x, y, t) = G_t * I_0(x, y, t)\,. \tag{1}$$

For each point $\vec{v}$ in the image sequence, we calculate the Shannon entropy of the signal histogram in a cylindrical neighborhood $N_s(\vec{s}, \vec{v})$ around it. That is,

$$H_D(s, d, \vec{v}) = -\sum_{q \in D} p(q, s, d, \vec{v}) \log p(q, s, d, \vec{v})\,, \tag{2}$$

The scale set for which the entropy is peaked is given by:

$$\begin{aligned} \hat{S}_p \quad = \quad & \{(s, d) : H_D(s-1, d, \vec{v}) < H_D(s, d, \vec{v}) > H_D(s+1, d, \vec{v}) \wedge \\ & H_D(s, d-1, \vec{v}) < H_D(s, d, \vec{v}) > H_D(s, d+1, \vec{v})\}\,, \end{aligned} \tag{3}$$

The saliency metric at the candidate scales is given by:

$$y_D(s,d,\vec{v}) = H_D(s,d,\vec{v}) W_D(s,d,\vec{v}), \quad \forall\,(s,d) \in \hat{S}_p\,, \tag{4}$$

The first term of eq. 4 is a measure of the variation in the information content of the signal. The weighting function $W_D(s,\vec{v})$ is a measure of how prominent the local maximum is at s, and is given by:

$$W_D(s,d,\vec{v}) = \frac{s^2}{2s-1} \sum_{q\in D} |p(q,s,d,\vec{v}) - p(q,s-1,d,\vec{v})| + d \sum_{q\in D} |p(q,s,d,\vec{v}) - p(q,s,d-1,\vec{v})|, \forall\,(s,d) \in \hat{S}_p\,, \tag{5}$$

where the values in front of each summation in the right part of eq. 5 are normalization factors. When a peak in the entropy for a specific scale is distinct, then the corresponding pixel probability density functions at the neighboring scales will differ substantially, giving a large value to the summations of eq. 5 and thus, to the corresponding weight value assigned. On the contrary, when the peak is smoother, then the summations in eq. 5, and therefore the corresponding weight, will have a smaller value. Let us note that we considered cylindrical neighborhoods of radius s and depth d for simplicity reasons. However, more complicated shapes, such as elliptical neighborhoods at different orientations and axes ratios could be considered.

2.2. Salient Regions

The analysis of the previous section leads to a set of candidate spatiotemporal salient points $S = \{(\vec{s}_i, \vec{v}_i, y_{D,i})\}$, where $\vec{v}_i = (x,y,t)$, $\vec{s}_i$ and $y_{D,i}$ are respectively, the position vector, the scale and the saliency value of the feature point with index i. In order to achieve robustness against noise we follow a similar approach as that in [9],[8] and develop a clustering algorithm, which we apply to the detected salient points. By this we define salient regions instead of salient points, the location of which should be more stable than the individual salient points, since noise is unlikely to affect all of the points within the region in the same way. The proposed algorithm removes salient points with low saliency and creates clusters that are a) well localized in space, time and scale, b) sufficiently salient and c) sufficiently distant from each other. The steps of the proposed algorithm can be summarized as follows:

1. Derive a new set S_T by applying a global threshold T to the saliency of the points that comprise S. Thresholding removes salient points with low saliency.
$$S_T = \{(\vec{s}_i, \vec{v}_i, y_{D,i}) : y_{D,i} > T\}\,. \tag{6}$$

2. Select the point i in S_T with the highest saliency value and use it as a seed to initialize a salient region R_k. Add nearby points j to the region R_k as long as the intra-cluster variance does not exceed a threshold T_V.
$$\frac{1}{|R_k|} \sum_{j\in R_k} c_j^2 < T_V\,, \tag{7}$$
where R_k is the set of the points in the current region k and c_j is the Euclidean distance of the jth point from the seed point i.

3. If the overall saliency of the region R_k is lower than a saliency threshold T_S, that is,
$$\sum_{j\in R_k} y_{D,j} \le T_S\,, \tag{8}$$
discard the points in the region back to the initial set of points and continue from step 2 with the next highest salient point. Otherwise, calculate the Euclidean distance of the center of region R_k from the center of salient regions already defined.

4. If the distance is lower than the average scale of R_k, discard the points in R_k back to the initial set, and continue from step 2 with the next highest salient point. Otherwise, accept R_k as a new cluster and store it as the mean scale and spatial location of the points in it.

5. Form a new set S_T consisting of the remaining salient points, increase the cluster index k and continue from step 2 with the next highest salient point.

To summarize, a new cluster is accepted only if it has sufficient local support, its overall saliency value is above the saliency threshold, and it is sufficiently distant in terms of Euclidean distance from already existing clusters.

3. Kernel-based Recognition of Human Actions

3.1. Distance Function

Using the feature detection scheme described in section 2, we represent a given image sequence by a set of features, where each feature corresponds to a cylindrical salient region of the image sequence in the space-time domain. A wide variety of classification schemes depend on the definition of an appropriate distance metric. We use the *Chamfer Distance* [2], as it can provide a distance measure between feature sets with unequal number of features. More specifically, for two feature sets $F = \{(x_i, y_i, t_i), 1 \le i \le M\}$ and $F' = \{(x'_j, y'_j, t'_j), 1 \le j \le M'\}$ consisting of an M and M' number of features, respectively, the Chamfer distance of the set F from the set F' is defined as follows:

$$D(F,F') = \frac{1}{M} \sum_{i=1}^{M} \min_{j=1}^{M'} \sqrt{(x'_j - x_i)^2 + (y'_j - y_i)^2 + (t'_j - t_i)^2}\,. \tag{9}$$

From eq. 9 it is obvious that the selected distance measure is not symmetrical, as $D(F,F') \neq D(F',F)$. For recognition purposes, it is desirable to select a distance measure that is symmetrical. A measure that satisfies this requirement is the average of $D(F,F')$ and $D(F',F)$, that is,

$$D_c(F,F') = \frac{1}{2}(D(F,F') + D(F',F))\,. \tag{10}$$

Let us note that for the calculation of the distance measure we only consider the spatiotemporal position of the detected salient points.

3.2. Space-Time Warping

There is a large amount of variability between feature sets due to differences in the execution speed of the corresponding actions. Furthermore, we need to compensate for possible shifting of the representations forward or backward in time, caused by imprecise segmentation of the corresponding actions. To cope with both these issues, we developed a linear space-time warping technique with which we model variations in time using a time-scaling parameter a and a time-shifting parameter b. In addition, in order to achieve invariance against scaling, we introduce a scaling parameter σ in the proposed warping technique. Prior to warping, we transform the spatial coordinates of the detected salient regions so that they have zero mean, in order to achieve invariance against translation. The parameters a, b, σ are estimated with a gradient-descent iterative scheme that minimizes the Chamfer distance between the sets. More specifically, let us denote by $F_w = \{(\sigma x_i, \sigma y_i, a \cdot t_i - b), 1 \leq i \leq M\}$ the feature set F with respect to feature set F'. Then, the distance between F' and F_w is given by eq. 9 as:

$$D(F_w, F') = \frac{1}{M} \sum_{i=1}^{M} \min_{j=1}^{M'} \sqrt{(x'_j - \sigma x_i)^2 + (y'_j - \sigma y_i)^2 + (t'_j - a \cdot t_i + b)^2} \,. \tag{11}$$

Similarly, the feature set F' with respect to feature set F can be represented as $F'_w = \{(\frac{1}{\sigma} x'_j, \frac{1}{\sigma} y'_j, \frac{1}{a} \cdot t'_j + b), 1 \leq j \leq M'\}$ and their distance as:

$$D(F'_w, F) = \frac{1}{M'} \sum_{j=1}^{M'} \min_{i=1}^{M} \sqrt{(x_i - \frac{1}{\sigma} x'_j)^2 + (y_i - \frac{1}{\sigma} y'_j)^2 + (t_i - \frac{1}{a} \cdot t'_j - b)^2} \,. \tag{12}$$

The distance to be optimized follows from the substitution of eq. 11 and eq. 12 to eq. 10. We follow an iterative gradient descent approach for the adjustment of the a, b and σ parameters. The update rules are given by:

$$a^{n+1} = a^n - \lambda_1 \frac{\partial D_c}{\partial a^n}, b^{n+1} = b^n - \lambda_2 \frac{\partial D_c}{\partial b^n}, \sigma^{n+1} = \sigma^n - \lambda_3 \frac{\partial D_c}{\partial \sigma^n} \,, \tag{13}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the learning rates and n is the iteration index. The iterative procedure stops when the values of a, b and σ do not change significantly or after a fixed number of iterations.

3.3. Relevance Vector Machine Classifier

A Relevance Vector Machine (RVM) [20] is a probabilistic sparse kernel model identical in functional form to the Support Vector Machines (SVM). In their simplest form, Relevance Vector Machines attempt to find a hyperplane defined as a weighted combination of a few Relevance Vectors that separate samples of two different classes. Given a dataset of N input-target pairs $\{(F_n, l_n), 1 \leq n \leq N\}$, an RVM learns functional mappings of the form:

$$y(F) = \sum_{n=1}^{N} w_n K(F, F_n) + w_0 \,, \tag{14}$$

where $\{w_n\}$ are the model weights and $K(.,.)$ is a Kernel function. Gaussian or Radial Basis Functions have been extensively used as kernels in RVM. In our case, we use as a kernel a Gaussian Radial Basis Function defined by the distance function of eq. 10. That is,

$$K(F, F_n) = e^{-\frac{D_c(F, F_n)^2}{2\eta}} \,, \tag{15}$$

where η is the Kernel width. RVM performs classification by predicting the posterior probability of class membership given the input F. In the two class problem, a sample F is classified to the class $l \in [0, 1]$, that maximizes the conditional probability $p(l|F)$. For L different classes, L different classifiers are trained and a given example F is classified to the class for which the conditional distribution $p_i(l|F)$,$1 \leq i \leq L$ is maximized, that is:

$$Class(F) = \arg\max_i (p_i(l|F)) \,. \tag{16}$$

4. Experimental Results

Similar to [1], we use aerobic exercises as a test domain. We created our own set of examples, consisting of 19 aerobic exercises, performed twice by four different subjects, leading to 152 corresponding feature sets. In Fig. 1 the salient regions detected in four instances of four sample image sequences are presented. The first two columns depict two executions of the same exercise by two different subjects while the last two columns depict the execution of another exercise by the same pair of subjects. It is apparent that there is consistency in the location and scale of the detected salient regions between different executions of the same exercise. The detected salient points seem to appear in areas with significant amount of activity, such as the points at which the hands move. Moreover, it seems that the scale of the detected regions is large when motion is fast (t_4, t'_2, t'_3, t'_4), and smaller when motion is slower (t_1, t_2, t_3, t'_1). Finally, the algorithm does not guarantee that detection of corresponding regions across the examples will occur at the same time instance. For example, at the time instances t_2 and t_3 in Fig. 1, the detection of the arms occurs at neighboring time instances.

In order to test the influence of our space-time warping algorithm, we randomly selected some examples from our original example set and we resized them to 1.2, 1.4, 1.6 and 1.8 times their initial size. We applied in these sequences the salient point detector of section 2 and we used the resulting representations in order to warp them in space and time with the original sequences. The result for a single pair

of original-resized sequences is shown in Fig. 2, where the latter has been cropped for illustration purposes, to show the estimated scale. We also stretched the resized sequence in time, so that its duration matches that of the original one. From the figure it is clear, that the space-time warped sequence is closer to the original one. The σ parameter from eq. 11 and eq. 12, for the resized sequence in Fig. 2, was equal to 1.18, which is very close to the actual value of 1.2. The difference can be attributed to the different size of the subjects performing the action. The error variance of the estimated σ was small, and equal to $0.04, 0.07, 0.08$ and 0.08, for scaling factors $1.2, 1.4, 1.6$ and 1.8 respectively.

We applied a Relevance Vector Machine Classifier to the available feature sets to address the classification problem. We constructed 19 classifiers, one for each class, and we calculated for each test example F the conditional probability $p_i(l|F), 1 \leq i \leq 19$. We performed our experiments in the leave-one-subject-out manner. That is, for estimating each $p_i(l|F)$, an RVM is trained by leaving out the example F and all other instances of the same exercise, that were performed by the subject from F. Each example was assigned to the class for which the corresponding classifier provided the maximum conditional probability, as depicted in eq. 16. The corresponding recall and precision rates are given in Table 4, along with the ones achieved by a k-nearest neighbor classifier. As can be seen from this Table, for many classes the recognition rate is pretty high, while for other classes it is much lower (e.g. classes 7, 13). In Table 4 the confusion matrix generated by the RVM classifier is given. It is obvious from this table that there are mutual confusions between specific classes, for instance classes 5, 6, 7, 12 and 13. The main differences between the corresponding actions are out of plane motions, not easily distinguished by a single camera. The global recall rate for the RVM classifier is 77.63%, an improved rate compared to [15], where spherical spatiotemporal neighborhoods were used. The performance achieved is relatively good, given the small number of examples with respect to the number of classes, and the fact that the subjects were not trained.

We used the average ranking percentile in order to measure the overall matching quality of our proposed algorithm. Let us denote by r^{F_n} the position of the correct match for the test example $F_n, n = 1 \ldots N_2$, in the ordered list of N_1 match values. Rank r^{F_n} ranges from $r = 1$ for a perfect match to $r = N_1$ for the worst possible match. Then, the average ranking percentile is calculated as:

$$\overline{r} = \left(\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{N_1 - r^{F_n}}{N_1 - 1} \right) 100\% \,. \qquad (17)$$

Our dataset consists of 152 image sequences divided in 19 classes, therefore $N_1 = 19$ and $N_2 = 152$. The average ranking percentile for the RVM classifier is 97.25%, which shows that for most of the missclassified examples, the correct matches are located in the first positions in the ordered list of match values.

Class Labels	1	2	3	4	5
kNN Recall/Prec.	1 / 1	1 / 1	1 / 1	1 / 1	0.38 / 0.5
RVM Recall/Prec.	1 / 1	1 / 1	1 / 1	0.75 / 1	0.38 / 0.6
Class Labels	**6**	**7**	**8**	**9**	**10**
kNN Recall/Prec.	0.25 / 0.67	0.63 / 0.45	1 / 1	0.75 / 1	1 / 0.8
RVM Recall/Prec.	0.25 / 0.5	0.63 / 0.63	0.75 / 1	0.75 / 1	1 / 1
Class Labels	**11**	**12**	**13**	**14**	**15**
kNN Recall/Prec.	0.75 / 1	0.75 / 0.75	0.38 / 0.21	1 / 1	0.5 / 0.29
RVM Recall/Prec.	0.88 / 1	0.5 / 0.57	0.75 / 0.35	1 / 0.89	0.88 / 0.64
Class Labels	**16**	**17**	**18**	**19**	**Total**
kNN Recall/Prec.	0 / 0	0.75 / 1	1 / 1	1 / 1	0.7434
RVM Recall/Prec.	0.5 / 0.5	1 / 0.89	0.88 / 0.88	0.88 / 0.78	0.7763

Table 1. Recall and Precision rates for the kNN and RVM classifiers

Figure 1. Detected spatiotemporal features in four sample image sequences, corresponding to two action classes, for four time instances, $t_i, t'_i, i = 1 \ldots 4$. For each class the detected regions are drawn for two different subjects performing the action.

Finally, we compared our method with past work on temporal templates[1]. In [1], each single-view test example was matched against seven views of each training example, performed several times by an experienced aerobics instructor. A performance of 66.67% (12 out of 18 moves) was achieved. Our training set consists of single view examples, performed several times by non expert subjects. Noise and shadow effects in our sequences create small, non-zero pixel regions in areas of the corresponding MEIs and MHIs where no motion exists. The recognition rate achieved is 46.71%. Removal of spurious areas with simple morphological operations led to deterioration in the overall performance.

5. Conclusions and Future Directions

In this paper, we extended the concept of saliency to the spatiotemporal domain in order to represent human motion with a sparse set of spatiotemporal features that, loosely speaking, correspond to activity peaks. We did this by measuring changes in the information content of neighboring pixels, in space and time. We devised an appropriate

Figure 2. Space-time warping. 1st column: Reference sequence, 2nd: space-time warped sequence, 3d: stretched sequence and 4th: original sequence

Class labels	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Total
1	**8**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
2	0	**8**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
3	0	0	**8**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
4	0	0	0	**6**	0	0	0	0	0	0	0	0	**1**	0	0	**1**	0	0		8
5	0	0	0	0	**3**	2	0	0	0	0	0	0	**3**	0	0	0	0	0	0	8
6	0	0	0	0	**2**	**2**	**2**	0	0	0	0	0	**2**	0	0	0	0	0	0	8
7	0	0	0	0	0	0	**5**	0	0	0	0	0	**1**	0	0	**1**	0	0	**1**	8
8	0	0	0	0	0	0	0	**6**	0	0	0	**2**	0	0	0	0	0	0	0	8
9	0	0	0	0	0	0	0	0	**6**	0	0	0	0	0	0	**1**	0	**1**	0	8
10	0	0	0	0	0	0	0	0	0	**8**	0	0	0	0	0	0	0	0	0	8
11	0	0	0	0	0	0	0	0	0	0	**7**	0	**1**	0	0	0	0	0	0	8
12	0	0	0	0	0	0	0	0	0	0	0	**4**	**4**	0	0	0	0	0	0	8
13	0	0	0	0	0	0	**1**	0	0	0	0	**1**	**6**	0	0	0	0	0	0	8
14	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	0	0	0	0	8
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**7**	**1**	0	0	0	8
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	**4**	0	0	0	8
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	0	8
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**7**	**1**	8
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	**7**	8
Total	8	8	8	6	5	4	8	6	6	8	7	7	17	9	11	8	9	8	9	152

Table 2. RVM Confusion Matrix

distance measure between sparse representations, based on the Chamfer distance, which allows us to use an advanced kernel-based classification scheme, the Relevance Vector Machines. We devised an iterative space-time warping technique which aligns in time the representations and achieves invariance against scaling, while translation invariance is achieved by appropriately transforming the features' location attributes. We illustrated the efficiency of our representation in recognizing human actions using as a test domain aerobic exercises. We presented results on real image sequences that illustrate the consistency in the spatiotemporal localization and scale selection of the proposed method. Finally, we presented comparative results with other methods in the literature, where the performance achieved by our proposed method was much higher.

In future research we wish to increase the power of the proposed method by investigating the extraction of spatiotemporal features around the detected salient points. This is a natural extension of similar methods that extract texture features around the detected points in the spatial domain. Making the method robust to rotation is also an important issue, which can be achieved by introducing an additional parameter for rotation in the proposed space-time warping technique.

Acknowledgements

The work presented in this article has been conducted at Delft University of Technology. It has been partially supported by the Netherlands BSIK-MultimediaN-N2 Interaction project. The work of A. Oikonomopoulos has been supported by the Greek State Scholaships Foundation (IKY). The data set was collected while I. Patras was with the ISIS group at the University of Amsterdam.

References

[1] A. Bobick and J. Davis. The Recognition of Human Movement Using Temporal Templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(3):257 – 267, 2001. 2, 4, 5

[2] G. Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(6):849–865, November 1988. 3

[3] L. Bretzner, I. Laptev, and T. Lindeberg. Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 405–410, 2002. 2

[4] S. Gilles. *Robust Description and Matching of Images*. PhD thesis, University of Oxford, 1998. 1

[5] R. Haralick and L. Shapiro. *Computer and Robot Vision II*. Addison-Wesley, 1993. Reading, MA. 1

[6] J. Hare and P. Lewis. Salient Regions for Query by Image Content. *International Conference on Image and Video Retrieval*, pages 317–325, 2004. 1

[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 1

[8] T. Kadir. *Scale, Saliency and Scene Description*. PhD thesis, University of Oxford, 2002. 1, 2, 3

[9] T. Kadir and M. Brady. Scale Saliency: A Novel Approach to Salient Feature and Scale Selection. *International Conference on Visual Information Engineering*, pages 25 – 28, 2000. 1, 2, 3

[10] T. Kirishima, K. Sato, and K. Chihara. Real-time gesture recognition by learning and selective control of visual interest points. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27:351 – 364, 2005. 2

[11] J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987. 1

[12] I. Laptev and T. Lindeberg. Space-time Interest Points. *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 432 – 439, 2003. 2

[13] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998. 1

[14] D. Lowe. Object Recognition from Local Scale-Invariant Features. *Proc. IEEE Int. Conf. Computer Vision*, 2:1150 – 1157, 1999. 1

[15] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition. *Proc. IEEE Int. Conference on Multimedia and Expo*, pages 430–433, 2005. 2, 5

[16] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997. 1

[17] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000. 1

[18] N. Sebe and M. Lew. Comparing salient point detectors. *Pattern Recognition Letters*, 24:89–96, 2003. 1

[19] B. ter Haar Romeny, L. Florack, A. Salden, and M. Viergever. Higher order differential structure of images. *Image and Vision Computing*, pages 317–325, 1994. 1

[20] M. Tipping. The Relevance Vector Machine. *Advances in Neural Information Processing Systems*, pages 652 – 658, 1999. 4

[21] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection in video. *IEEE International Conference on Systems, Man and Cybernetics*, pages 635–640, 2004. 2