SPATIOTEMPORAL SALIENCY FOR HUMAN ACTION RECOGNITION

A. Oikonomopoulos
Delft University of Technology
A.Oikonomopoulos@ewi.tudelft.nl

I. Patras
Delft University of Technology
I.Patras@ewi.tudelft.nl

M. Pantic
Delft University of Technology
M.Pantic@ewi.tudelft.nl

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 changes in the information content of pixel neighborhoods not only in space but also in time. We introduce an appropriate distance metric between two collections of spatiotemporal salient points that is based on the Chamfer distance and an iterative linear time warping technique that deals with time expansion or time compression issues. We propose a classification scheme that is based on Relevance Vector Machines and on the proposed distance measure. We present results on real image sequences from a small database depicting people performing 19 aerobic exercises.

1. INTRODUCTION

Recognition and interpretation of human activities is by itself a significant research area 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 arrive at a semantic description of the content of an image sequence we need not use all available information. A good description can be obtained by considering only the information around certain points of interest.

For content-based image retrieval applications, the notion of interesting points has been extensively used. According to Haralick and Shapiro [1] 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. [2] detect interesting points using a Harris corner detector and estimate gray value differential image invariants [3][4] at different scales. Loupias and Sebe [5] use wavelet-based salient point detectors in order to detect global and local variations in images for content-based image retrieval applications. Gilles introduces the notion of saliency in terms of local signal complexity or unpredictability in [6]. Interesting point detectors and local descriptors are compared in [7][8][9] in terms of repeatability rate and information content.

An important issue in salient point detection is automatic selection of the scale at which the salient points will be detected and at which local features will be extracted. Lindeberg et al. [10] integrate a scale-space approach for corner detection and search for local extremes across scales. Kadir and Brady [11] extend the original Gilles algorithm and estimate the information content 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 both the entropy and its rate of change at the scale in question. In [12], the performance of the salient point detector developed in [11] and an object recognition approach using keypoints developed by Lowe in [13] is examined.

While a large amount of work has been done on image-based retrieval and object recognition, the concept of saliency has only recently begun to be used for space-time content-based video retrieval and for activity recognition. In [14], 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, such as stopping or starting. In [15], an input image sequence is 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 spatiotemporal image and a Mahalanobis distance metric is applied between the sets in order to discriminate different activities.

In this paper, we detect spatiotemporal features in given image sequences by extending in the temporal direction the information-theoretic salient feature detector developed in [11]. In contrast to the work of Laptev [14], in which a sequence is represented by the local activity endpoints (starts/stops), our representation contains the spatiotemporal points at which there are peaks in activity variation such as the edges of a moving object. We use the Chamfer distance as an appropriate distance metric between two representations and we propose a linear time warping technique in order to deal with different speeds in the execution of the actions. A simple kNN classifier and one based on Relevance Vector Machines, introduced in [16], are used in order to test the efficiency of the representation. We test the proposed method using real image sequences and we present experimental results which 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 time warping technique. In section 4, we present our experimental results, and in section 5, final conclusions are drawn.

2. FEATURE DETECTION

2.1. Spatiotemporal saliency

Let us define $N_c(s, \vec{v})$ as the circular neighborhood of pixels of radius s in an image I, centered at $\vec{v} = (x, y)$. In [11], Kadir and Brady define a saliency metric y_D in order to detect salient points in static images, the calculation of which is done by measuring changes in the information content of N_c for a set of different radiuses (i.e scales). In order to detect spatiotemporal salient points

0-7803-9332-5/05/$20.00 ©2005 IEEE

at peaks of the activity variation, we extend this approach by defining $N_s(s,\vec{v})$ as the spherical neighborhood of pixels of radius s in a given image sequence, centered at the spatiotemporal point $\vec{v} = (x,y,t)$ and we define the spatiotemporal saliency $y_D(s,\vec{v})$ by measuring the changes in the information content within $N_s(s,\vec{v})$. We consider as input signal the convolution of the intensity information with a first-order Gaussian filter. Gaussian-derivative filters have been extensively used for detecting interesting points in static images. Here, we apply them in the temporal domain in order to arrive at a measure of activity. For each point $\vec{v} = (x,y,t)$ in an image sequence, the Shannon entropy of the intensity histogram in a spherical region of radius s around the point is defined by:

$$H_D(s,\vec{v}) = -\int\limits_{q\in D} p_D(s,\vec{v}) \log_2 p_D(s,\vec{v}) dq, \quad (1)$$

where $p_D(s,\vec{v})$ is the probability density of the signal histogram as a function of scale s and position $\vec{v}$. With q we denote the signal value and with D the set of all signal values. Let us define as $\vec{s}$ the vector of scales at which the entropy is peaked, that is,

$$\vec{s} = \left\{ s | \frac{\partial H_D(s,\vec{v})}{\partial s} = 0 \bigwedge \frac{\partial^2 H_D(s,\vec{v})}{\partial s^2} < 0 \right\}. \quad (2)$$

Then, following the approach defined at [11] we can define the saliency metric at the candidate scales as follows:

$$y_D(\vec{s},\vec{v}) = H_D(\vec{s},\vec{v}) \times W_D(\vec{s},\vec{v}), \quad (3)$$

Eq. 3 gives a measure of how salient a point $\vec{v}$ is at certain scales $\vec{s}$, where we consider only the scales at which the local entropy in the spherical pixel neighborhood around it is locally maximized. The first term of eq. 3 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 scale s, and is given by:

$$W_D(s,\vec{v}) = \frac{s^2}{s-1} \int\limits_{q\in D} \left| \frac{\partial}{\partial s} p_D(s,\vec{v}) \right| dq. \quad (4)$$

2.2. Salient Regions

The analysis of the previous subsection leads to a set of candidate spatiotemporal salient points $S = \{(s_i, \vec{v}_i, y_{D,i})\}$, where $\vec{v}_i = (x,y,t)$, $s_i, y_{D,i}$ are respectively, the position vector, the scale and the saliency value of the feature point with index i. In order to make the feature detector more robust against noise we developed a clustering algorithm, which we apply to S. By clustering S, we wish to remove salient points with low saliency value and create clusters that are well localized in space and time and sufficiently distant from each other. In addition, we want to take the saliency of the points into consideration such that the overall saliency of the region is above a fixed threshold, and regions with overall saliency lower than the threshold are discarded, as not salient enough. The steps of the proposed algorithm can be summarized as follows :

1. Remove salient points with saliency value below a global threshold T and derive a new set S_T from S, that is,
$$S_T = \{(s_i, \vec{v}_i, y_{D,i}) : y_{D,i} > T\}.$$
2. Select the point with index i in S_T that has the highest saliency value. Use the salient point i as a seed to initialize a salient region R_k (in the first iteration k = 1).
3. Add nearby points j to the region R_k as long as the within cluster variance does not exceed a threshold V_{th}. That is, as long as:
$$\frac{1}{|R_k|} \sum_{j\in R_k} d_j^2 < V_{th},$$
where R_k is the set of points in the current region k and d_j is the Euclidean distance of the jth point from seed point i.
4. If the overall saliency of the region is lower than a saliency threshold S_{th}, that is,
$$\sum_{j\in R_k} y_{D,j} \leq S_{th}$$
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 Euclidian distance of the center of region R_k from the center of salient regions already defined, that is, from salient regions $R_{k'}, k' < k$.
5. If the distance is lower than the average scale of the region, discard the points in the region back to the initial set of points and repeat from step 2 with the next highest salient point. Otherwise, accept the region as a new cluster and store as the mean scale and spatial location of the points contained in it.
6. Form a new set S_T consisting of the remaining salient points and continue from step 2 with the next highest salient point.

In order to increase its execution speed, we apply the proposed algorithm in two levels. At *Level 1* we apply the algorithm for each frame, by considering only circular pixel neighborhoods. We derive in this way feature sets of the form $F_t = \{(x_{t,i}, y_{t,i}, s_{t,i}), 1 \leq t \leq K, 1 \leq i \leq L_t\}$, where K is the total number of frames and L_t is the total number of salient regions detected in frame t. At *Level 2* we consider only the locations of the salient regions detected in the first level and we apply the algorithm for spherical pixel neighborhoods, thus by taking into account neighboring frames as well. We derive in this way a feature set of the form $F' = \{(x'_j, y'_j, t'_j, s'_j), 1 \leq j \leq L\}$, where L is the total number of the detected salient regions. By applying the algorithm in two steps, we discard from the computations of the second step image points that are not salient in space and therefore not salient in time.

3. RECOGNITION

A wide variety of classification schemes, ranging from kNN to Support Vector Machines, depends on the definition of an appropriate distance metric. We use the *Chamfer Distance*, as it provides a distance measure between feature sets with unequal number of features. For two feature sets $F = \{(x_i, y_i, t_i), 1 \leq i \leq M\}$ and $F' = \{(x'_j, y'_j, t'_j), 1 \leq j \leq M'\}$ 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:

$$D(F,F') = \frac{1}{M} \sum_{i=1}^{M} \min_{j=1}^{M'} \sqrt{(x'_j - x_i)^2 + (y'_j - y_i)^2 + (t'_j - t_i)^2}. \quad (5)$$

From eq. 5 it is obvious that the selected distance metric is not symmetrical, as $D(F,F') \neq D(F',F)$. For recognition purposes, it is desirable to select a distance metric that is symmetrical. A metric that satisfies this requirement is the average of $D(F,F')$ and $D(F',F)$, that is,

$$D_c(F,F') = \frac{1}{2}(D(F,F') + D(F',F)). \quad (6)$$

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

3.1. Time Warping

Differences in the execution speed of similar or different actions performed by the same or different subject, as well as possible shifting in time, makes it impossible to compare corresponding feature sets. We propose a linear time warping technique in order to cope with both these issues, by introducing a time-scaling parameter α and a time-shifting parameter β. With the proposed model, time warping becomes a simple optimization problem, in which we try to minimize the Chamfer distance between two feature sets by adjusting the values of the α, β parameters. By warping feature set F to feature set F' we impose their Chamfer distance to become equal to:

$$D(F_w, F') = \frac{1}{M}\sum_{i=1}^{M}\min_{j=1}^{M'}\sqrt{(x'_j - x_i)^2 + (y'_j - y_i)^2 + (t'_j - \alpha \cdot t_i + \beta)^2}. \quad (7)$$

Similarly, by warping feature set F' to feature set F their Chamfer distance becomes:

$$D(F'_w, F) = \frac{1}{M'}\sum_{j=1}^{M'}\min_{i=1}^{M}\sqrt{(x_i - x'_j)^2 + (y_i - y'_j)^2 + (t_i - \frac{1}{\alpha} \cdot t'_j - \beta)^2}. \quad (8)$$

The distance to be optimized follows from the substitution of eq.7 and 8 to eq. 6. We follow a gradient descent approach for the adjustment of the α, β parameters. The update rules are given by:

$$\alpha^{n+1} = \alpha^n - \lambda_1 \frac{\partial D_c}{\partial \alpha^n}, \beta^{n+1} = \beta^n - \lambda_2 \frac{\partial D_c}{\partial \beta^n}, \quad (9)$$

where λ_1 and λ_2 are the learning rates and n is the iteration index. The algorithm terminates after a fixed number of iterations, or after the values of α and β do not significantly change.

3.2. Classification

We propose a classification scheme based on Relevance Vector Machines in order to classify given examples of human actions. Predictions in RVM are probabilistic, in contrast with the hard decisions provided by SVM. 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:

$$y(F) = \sum_{n=1}^{N} w_n K(F, F_n) + w_0, \quad (10)$$

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 and can be viewed as a distance metric between F and F_n. In our case, we use as a kernel the distance function defined in eq. 6. RVM performs classification by predicting the posterior probability of class membership given the input F. The conditional probability $P(l_n|w, F_n)$ is given by:

$$P(l_n|w, F_n) = \sigma\{y(F_n)\}^{l_n}[1 - \sigma\{y(F_n)\}]^{1-l_n}, \quad (11)$$

where $\sigma(y) = 1/(1+e^{-y})$ is the logistic sigmoid function. 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:

$$Class(F) = \arg\max_i(p_i(l|F)). \quad (12)$$

4. EXPERIMENTAL EVALUATION

For our experiments, we used the same set of aerobic exercises as Bobick and Davis [15], but performed by different subjects. Our dataset consists of 152 test image sequences forming 19 different motion classes, that is, eight examples per class, performed by four different subjects. In Fig. 1 the salient regions detected in six instances of four sample image sequences are presented. It is apparent that there is consistency in the location and scale of the detected spatiotemporal salient regions between different executions of the same exercise. The detected salient points seem to correspond at peaks of activity, such as the points in space and time at which the hands move fast. Moreover, there seems to be a correlation between the scale of the detected regions and the motion magnitude, that is, the scale of the detected regions is large when the motion is fast (instances $t_2, t_3, t_4, t'_2 \ldots t'_5$), and smaller when the motion is slower ($t_1, t_5, t_6, t'_1, t'_6$). This can be explained by the fact that when the motion is fast, the activity spreads over a larger spatial region than to when the motion is slow. Furthermore, it is apparent that although the spatiotemporal localization of the salient points is good, the algorithm does not try to guarantee detection of the same number of regions at a specific time instant. For example, at the time instances t'_3 and t'_4 of the second pair of image sequences of Fig. 1 (i.e. last two columns), the detection of the head does not occur at the same, but at neighboring time instances.

We apply a classifier based on Relevance Vector Machines in order to test the efficiency of the representation. We trained 19 RVM classifiers, one for each class and we calculated for each example F_n the conditional probability $p_i(l_n|F_n), 1 \leq i \leq 19, 1 \leq n \leq 152$. Each example was assigned to the class for which the corresponding classifier provided the maximum conditional probability, according to eq. 12. For estimating each of the $p_i(l_n|F_n)$, an RVM is trained by leaving out the example F_n as well as all other instances of the correct class that were performed by the subject that performed F_n. We compared the performance of the RVM approach with that of a simple 1 nearest neighbor classifier. The corresponding recall and precision rates are given in Table 1. We notice a considerable improvement in the recall and precision rates for most of the action classes by using an RVM classifier, leading to an increase of almost 8% in the global recognition rate.

We used the average ranking percentile in order to measure the matching quality of our proposed algorithm. Let rank r^{F_n} denote the position of the correct match for test example $F_n, n = 1 \ldots 152$, in the ordered list of 19 match values, provided for each example by the 19 trained RVM classifiers. Rank r^{F_n} ranges from $r = 1$ for a perfect match to $r = 19$ for the worst possible match. Then, the average ranking percentile is calculated as follows:

$$\overline{r} = \left(\frac{1}{152}\sum_{i=1}^{152}\frac{19 - r^{F_n}}{19 - 1}\right)100\% \quad (13)$$

The average rank percentile for the RVM classifier was calculated equal to 97.15%. Its high value denotes that the majority of the correct matches for the missclassified examples are located in the first positions in the ordered list of match values.

5. SUMMARY AND CONCLUSIONS

In this paper, we extended the concept of saliency in the spatiotemporal domain, in order to represent human motion by using a sparse set of spatiotemporal features that, loosely speaking, correspond to activity peaks. We did this by measuring changes in

Fig. 1. Detected spatiotemporal features in four sample image sequences, corresponding to two action classes, for six time instances, $t_i, t_i', i = 1 \ldots 6$.

the information content of neighboring pixels, not only in space but also in time. We devised an appropriate distance measure between sparse representations containing different numbers of features based on the Chamfer distance. The proposed distance measure allows us to use an advanced kernel-based classification scheme, the Relevance Vector Machine. Our results on real image sequences illustrate the consistency of the proposed method in the spatiotemporal localization and scale selection. Furthermore, the classification results clearly illustrate the superiority of the proposed kernel-based classification scheme over the simple kNN classification.

In future research we aim to increase the discriminating power by investigating the extraction of spatiotemporal features around the spatiotemporal salient points. This, will come as a natural extention of similar methods that in the spatial domain extract texture features around the detected spatial points. An issue that also has to be considered is finding better clustering techniques, which can enhance the efficiency of the representation.

Acknowledgements

The work of A.Oikonomopoulos is supported by the Greek State Scholaships Foundation (IKY). The work of I. Patras and M. Pantic is supported by the Netherlands BSIK-MultimediaN-N2 Interaction project. The data set was collected while I.Patras was with the ISIS group at the University of Amsterdam.

6. REFERENCES

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

Class Labels	1	2	3	4	5	6	7	8	9	10
kNN Recall	1	0.88	1	0.88	0.13	0.75	0	0.88	1	0.5
kNN Precision	1	1	1	0.88	0.5	0.67	0	0.88	1	1
RVM Recall	1	1	1	0.88	0.5	0.75	0.25	0.88	1	1
RVM Precision	0.89	1	0.89	0.88	0.67	0.6	0.29	0.88	1	0.89

Class Labels	11	12	13	14	15	16	17	18	19	Total
kNN Recall	0.13	0.75	0	0.88	0.75	0.63	0.75	1	0.75	0.6645
kNN Precision	0.5	0.6	0	1	0.5	0.83	0.86	0.67	0.22	0.6645
RVM Recall	1	0.5	0.38	1	0.5	0.5	0.63	0.88	0.5	0.7434
RVM Precision	1	0.4	0.33	1	0.44	0.5	0.56	1	0.67	0.7434

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

[2] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–535, May 1997.

[3] J.J. Koenderink and A.J. van Doorn, "Representation of local geometry in the visual system," *Biological Cybernetics*, vol. 55, pp. 367–375, 1987.

[4] B.M. ter Haar Romeny, L.M.J. Florack, A.H. Salden, and M.A. Viergever, "Higher order differential structure of images," *Image and Vision Computing*, pp. 317–325, July/August 1994.

[5] E. Loupias, N. Sebe, S. Bres, and J.-M. Jolion, "Wavelet-based salient points for image retrieval," *Proc. IEEE Int. Conference on Image Processing*, vol. 2, pp. 518 – 521, September 2000.

[6] S. Gilles, *Robust Description and Matching of Images*, Ph.D. thesis, University of Oxford, 1998.

[7] N. Sebe and M.S. Lew, "Comparing salient point detectors," *Pattern Recognition Letters*, vol. 24, pp. 89–96, 2003.

[8] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, June 2000.

[9] C. Schmid and K. Mikolajczyk, "A performance evaluation of local descriptors," *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 257–263, June 2003.

[10] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 77–116, Nov. 1998.

[11] T. Kadir and M. Brady, "Scale saliency: a novel approach to salient feature and scale selection," *Int. Conf. on Visual Information Engineering*, pp. 25 – 28, November 2000.

[12] J.S. Hare and P.H. Lewis, "Salient Regions for Query by Image Content," *Int. Conf. on Image and Video Retrieval*, pp. 317–325, July 2004.

[13] D.G. Lowe, "Object recognition from local scale-invariant features," *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, pp. 1150 – 1157, September 1999.

[14] I. Laptev and T. Lindeberg, "Space-time Interest Points," *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 432 – 439, October 2003.

[15] A.F. Bobick and J.W. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257 – 267, March 2001.

[16] M.E.Tipping, "The Relevance Vector Machine," *Advances in Neural Information Processing Systems*, pp. 652 – 658, September 1999.