Recurrence Matrices for Human Action Recognition

V. Javier Traver[1,2], Pau Agustí[1,2] and Filiberto Pla[1,2]
[1]*DLSI, Jaume-I University, Castellón, Spain*
[2]*iNIT, Jaume-I University, Castellón, Spain*

Keywords: Human Action Recognition, Recurrence Matrices, Frame Descriptors, Motion and Shape Cues, Action Characterization, Temporal Information.

Abstract: One important issue for action characterization consists of properly capturing temporally related information. In this work, recurrence matrices are explored as a way to represent action sequences. A recurrence matrix (RM) encodes all pair-wise comparisons of the frame-level descriptors. By its nature, a recurrence matrix can be regarded as a temporally holistic action representation, but it can hardly be used directly and some descriptor is therefore required to compactly summarize its contents. Two simple RM-level descriptors computed from a given recurrence matrix are proposed. A general procedure to combine a set of RM-level descriptors is presented. This procedure relies on a combination of early and late fusion strategies. Recognition performances indicate the proposed descriptors are competitive provided that enough training examples are available. One important finding is the significant impact on performance of both, which feature subsets are selected, and how they are combined, an issue which is generally overlooked.

1 INTRODUCTION

Human action recognition has been receiving remarkable research effort for about two decades (Aggarwal and Ryoo, 2011) due to both the difficulty of the problem and its wide range of applications such as visual surveillance, human-machine interaction, or team sports analysis, to name but a few.

One of the relevant issues for action representation is to properly capture the temporal information. Some solutions involve accumulating local histograms along time (Lucena et al., 2012), extract a short-time series of a few still snapshots of representative poses (Brendel and Todorovic, 2010), or decompose actions into sequences of "actoms" (key atomic action units), and weight visual features by their temporal distance to these actoms (Gaidon et al., 2011a). There is some recent interest in enriching the popular bag-of-words representation with temporal information (Matikainen et al., 2010). Considering multiple temporal scales (Niebles et al., 2010) can be effective for modelling human activity which may include simpler and shorter actions.

In this work, feature vectors describing the action at frame level at two different time steps in the image sequence are compared, thus producing a 2D *recurrence matrix* (RM) of pair-wise distances which captures all the frame-to-frame similarities, and it can therefore be viewed as a time-holistic representation providing a rich characterization of the temporal structure and evolution of the action. However, this information implicitly contained in the recurrence matrices have to be summarized in the form of an appropriate RM descriptor, which is finally used for learning and recognizing actions.

This or similar representations were used in (Cutler and Davis, 2000) to analyse periodic motion and distinguish running bipeds (humans) from quadrupeds (dogs), and in (BenAbdelkader et al., 2004) for gait-based biometrics. A related approach, a delay-embedding technique, was developed in (Ali et al., 2007) for action recognition from trajectories of body landmarks. An auto-correlation kernel for time series has been recently proposed for action recognition (Gaidon et al., 2011b). The most related work is (Junejo et al., 2011), where temporal self-similarity matrices are explored as an appropriate view-invariant action representation. Our focus is not on view invariance, but in exploring alternative representations both to build the recurrence (or self-similarity) matrices, and to derive the RM descriptor. Additionally, a general procedure to combine RM-level descriptors using a combination of early and late fusion strategies is investigated.

Traver V., Agustí P. and Pla F..
Recurrence Matrices for Human Action Recognition.
DOI: 10.5220/0004216202710276
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 271-276
ISBN: 978-989-8565-47-1

2 METHODOLOGY

The proposed system consists of a *frame descriptor* (Section 2.1) that describes every frame of a given input action sequence, a *recurrence matrix* (RM) (Section 2.2) that compares all frame descriptors pairwise, and an *RM descriptor* (Section 2.3) that summarizes the RM to characterize the action category. One or several RM descriptors are finally used to represent any action sequence. Learning action categories and recognizing new (unseen) action instances rely on these RM descriptors. To be more precise, the RM descriptors corresponding to a single action sequence are combined using both early and late fusion (Section 2.4).

2.1 Frame Descriptor

The Tran-Sorokin descriptor (Tran and Sorokin, 2008) is used in this work to characterize the actions at individual frames within an action sequence. This is descriptor includes both motion and shape visual cues (Table 1) into the "single-frame descriptor" (SFD) comprising 216 features. The SFD of three 5-frame time windows, each corresponding to current frames ($[t-2,t+2]$), past frames ($[t-7,t-3]$), and future frames ($[t+3,t+7]$), are separately concatenated and PCA-projected to reduce the overall dimensionality. These parts (CFW, PFW and FFW) provide temporally contextual information which enrich the representation at a single frame.

2.2 Recurrence Matrix

Let $\mathbf{f}_t$ be a frame descriptor at a discrete time $t \in \{1,\ldots,T\}$, for a video sequence of T frames. Then, the recurrence matrix $\mathbf{R}$ is computed from pairwise distances of the frame descriptors, $\mathbf{R}(i,j) = d(\mathbf{f}_i,\mathbf{f}_j),\ i,j \in \{1,\ldots,T\}$. A binary version of this matrix can be obtained by thresholding distances: $\mathbf{R}_\theta(i,j) = H(d(\mathbf{f}_i,\mathbf{f}_j)-\theta)$, where H is the Heaviside step function: $H(x) = 1$ if $x > 0$ and $H(x) = 0$ otherwise.

The proposed recurrence matrix representation is inspired by the idea of recurrence plots (Marwan et al., 2007). Recurrence plots allow to visually analyse or automatically quantify the properties or behaviour of dynamical systems. Although these plots may be computed using concepts of phase space and time delay methods, in this work we just used the concepts of state and state-to-state comparison. We consider the action sequence as the dynamical system, and the state at a given time is the snapshot of the action at that time. Such an state is represented with a chosen frame descriptor.

For the distance function d, the Euclidean distance normalized by the length of the frame descriptor was chosen. This normalization aims at removing the effect of the length of different frame descriptors.

2.3 Describing a Recurrence Matrix

Given a recurrence matrix, a description of it is required as the final representation of the underlying action. We have explored two different RM descriptors: the histogram of line lengths (HoL) and the projections along anti-diagonals (PaD).

Histogram of Line Lengths (HoL). Some measures proposed for the recurrence quantification analysis (RQA) are based on the diagonal lines of a binary recurrence matrix. A diagonal of length l in our recurrence matrix means that the action is similar during l frames, according to the frame descriptor, distance function, and threshold used. Instead of using individual measures derived from these diagonals, such as entropy or determinism (Marwan et al., 2007), we propose to use an histogram of the diagonal lengths for a set of lengths $\mathcal{L} = \{l_1,l_2,\ldots,l_n\}$, which is expected to provide a richer representation than a few individual measures. To further enrich the descriptor, the histogram of vertical lines lengths is also considered. Both histograms are separately normalized. To find the lengths of the vertical and diagonal lines, the CRP Toolbox[1] (Marwan et al., 2007) is used. In the experiments reported below, we used $\mathcal{L} = \{1,2,\ldots,20\}$.

Therefore, the histogram of (diagonal and vertical) line lengths (HoL) has some useful properties for our problem. It provides a *global* quantification of action dynamics, and it is not sensitive to the starting point of the action. Being an histogram, it is also robust to whether the action has cycles or repetitions and, to some extent, to the speed of the performance of the action as well.

Before thresholding, we smooth the RM with a Gaussian filter ($\sigma = 5$) on 5×5 local windows. This choice was somehow arbitrary and subject to further experimentation. Since an appropriate threshold is not easy to choose, we obtain several binary recurrence matrices with different thresholds. In the experiments reported here, we experimented with three thresholds expressed as the 30%, 40% and 50% quantiles of the contents of the binary RM. Further work is needed to explore the effect of these thresholds or whether a single threshold suffices and can be learned.

[1] http://tocsy.pik-potsdam.de/CRPtoolbox

Table 1: Parts of the Tran-Sorokin descriptor (Tran and Sorokin, 2008). Font styles, '+' symbols and indentation are given to denote subparts.

Short name	Description	Number of features
MCD	Motion Context Descriptor, full descriptor	**286**
SFD	Single-Frame Descriptor (OF+SH)	216
+ OF	Optical flow, (local) motion	+144
+ SH	Silhouette, shape	+72
CONTX	Temporally contextual information	70
+ CFW	Current-frame window, present	+50
+ PFW	Past-frame window, past	+10
+ FFW	Future-frame window, future	+10

Projections along anti-Diagonals (PaD). Areas in the RM which are close to the main diagonal represent the temporally local information, which can be captured and summarized by projecting the information in the RM along the lines orthogonal to the main diagonal. We refer to these lines as the anti-diagonals. Each projection is computed by a weighed sum of the anti-diagonals, where the weight decays away from the main diagonal. The resulting projection is then normalized and resized to a fixed size w (we set $w = 40$) so that sequences of different lengths have feature vectors of the same size.

Unlike HoL, the Projections Along anti-Diagonals (PaD) better captures the temporally *local* information, and it is sensitive to the starting point and the number of cycles of the action. Another difference is that HoL requires the RM to be binarized to be able to compute the lengths of the diagonal lines, whereas PaD can be computed both with unthresholded and binary RMs. Here, we use unthresholded RMs.

In both, HoL and PaD descriptors, for the SH and OF parts of the descriptor, the features are split according to the 2×2 grid considered in the bounding box in (Tran and Sorokin, 2008). Then, independent recurrence matrices and descriptors are computed for each of the 4 cells in the grid, as well as for the global feature set. As a result, five RM descriptors sets result from any of the SH or OF parts.

2.4 Feature Combination

One procedure to build the final RM-based action descriptor could consider the full 286-feature Tran-Sorokin descriptor as the frame descriptor, then build the RM, and finally describe the RM by using HoL or PaD. However, we explore a more general procedure where arbitrary parts of the Tran-Sorokin descriptor can be used and combined flexibly, so that the different roles played by motion, shape and temporally contextual cues can be studied. Additionally, having several RMs, each computed from separate pieces of visual information, can be more discriminative than having a single RM of the complete information taken as a whole (Serra-Toro and Traver, 2011).

The notation we follow to represent how the final features Φ are computed is $\Phi = \{\langle \mathbf{F}_1 \rangle, \ldots, \langle \mathbf{F}_N \rangle\}$, where $\mathbf{F}_i = \{\mathbf{f}_{i1} \ldots \mathbf{f}_{in_i}\}$, denotes a set of frame descriptors $\mathbf{f}_{ij}$, and $\langle \mathbf{F} \rangle$ represents the concatenation of the RM descriptors resulting from the recurrence matrices obtained from each of the frame descriptors in $\mathbf{F}$. Finally, the set of N concatenations of RM descriptors are used in separate classifiers and a max-score fusion scheme is adopted. Let us have a look at a couple of clarifying examples using parts of the Tran-Sorokin descriptor (Table 1) as frame descriptors: (1) $\{\langle \mathsf{sh} \rangle, \langle \mathsf{of} \rangle\}$ means that two sets of RM descriptors are built, one using the shape features and another one using the motion features; (2) $\{\langle \mathsf{sh} \rangle, \langle \mathsf{pfw}, \mathsf{ffw} \rangle\}$ involves also two sets of RM descriptors, but the second one is, in turn, a concatenation of the RM descriptors resulting from separately considering the PFW and FFW parts. Please note that $\langle \mathsf{pfw}, \mathsf{ffw} \rangle$ represents the concatenation of two RM descriptors, not *one* RM descriptor computed from the concatenation of the two frame descriptors, pfw and ffw.

Notice this representation is fairly general and combines *early* fusion (by concatenating RM descriptors —operator $\langle\rangle$) with *late* fusion (by combining the RM descriptors at the decision level —operator $\{\}$). In some contexts, this kind of combining early and late fusion is shown to be advantageous over using only early or late fusion separately (Lan et al., 2012).

Individual features of the final feature vector are normalized through standardization (i.e. normalized features have zero mean and unit variance).

Some Remarks. Due to its exploratory nature, this work does not aim at comparing the proposed approach with that of (Junejo et al., 2011). However, a rough comparison is provided in one part of the ex-

Table 2: Features of the used action datasets.

	Weizmann	UIUC1	IXMAS
No. actions	10	14	11
No. examples	90	532	1980
Mean sequence length (frames)	62	82	77

perimental section to assess the potential of the proposed descriptors and combination of features. To better contextualize this comparisoin, the main differences in both works follow. Junejo et al. (Junejo et al., 2011) mainly address the computation of the self-similarity matrix from point trajectories, but also consider image-based representations. We focus on an image-based representation. To that end, they use the HOG descriptor (Dalal and Triggs, 2005), which is a generic descriptor that captures shape/appearance information, and was originally proposed for pedestrian detection, whereas the Tran-Sorokin frame descriptor that we use, is a richer descriptor, and specifically proposed for action recognition. Indeed, for this reason, Junejo et al. also explore the optic flow information to complement the appearance cues in HOG.

For describing the self-similarity matrix (SSM), Junejo et al. use histograms of gradient orientations over log-polar grids centred on the main diagonal of unthresholded SSMs. Then, an orderless bag-of-words (BoW) representation of the resulting descriptor is used and each action sequence is finally represented as one histogram. In contrast, we use simpler descriptors, of both global (HoL) and semi-local (PaD) nature, and do not use any BoW representation.

3 RESULTS

We used the Weizmann (Gorelick et al., 2007), UIUC1 (Tran and Sorokin, 2008), and IXMAS (Weinland et al., 2006) datasets of human actions. The Tran-Sorokin descriptor computed for these datasets is publicly available (Tran-Sorokin, 2008). Some relevant features of these datasets are given in Table 2. For classification, the Nearest Neighbour (NN) and a linear Support Vector Machines (SVM) are used. For evaluation, the leaving-one-out (LOO) and leaving-one-actor-out (LOAO) protocols are used. When using the multi-view dataset IXMAX, some other protocols (described in an experiment below) are employed.

First, we evaluate the performance of three different parts of Tran-Sorokin descriptor as a frame descriptor, and the two proposed RM descriptors. Results on LOO on the three datasets (Table 3) suggests that HoL is a better descriptor in the Weizmann dataset, while PaD is generally better in UIUC1 and IXMAS datasets. Regarding the discriminative power of motion and shape cues, there is not a clear winner, although shape tends to offer better results than motion. An interesting observation is the role of the FFW part of the frame descriptor which, despite its low dimensionality (just 10 features), gives comparable or better results than shape and motion cues, which are about 7 and 14 times larger, respectively. This can be explained by the way the FFW is computed which integrates shape and motion features over a short-time window. Finally, the poor results on the Weizmann dataset can be due to the limited amount of training sequences available (9 examples per action) and that these sequences are also shorter than those in the other datasets.

Next, we compare different feature combinations. Results (Table 4) clearly indicates that combining features of different nature (e.g. optic flow and silhouette) tends to outperform the use of these features separately. This can be seen by comparing the LOO columns in Table 4 with the corresponding lines in Table 3 that use parts of the Tran-Sorokin descriptor separately. In addition, which features are used and how they are combined can make a significant difference. Notice, for instance, that Φ_3, Φ_4 and Φ_5 are three different ways of combining sh, of and ffw features, which result in different performances. While combination Φ_3 gives the best results in most cases (notice the column-wise best results, which are boldfaced), the optimum way of combining the features is data-dependent. This calls for an automatic procedure that efficiently chooses a feature combination which optimizes computational and recognition criteria. While it is commonly known the complementarity of different visual features for action recognition, not much work has been done on finding optimal ways of fusing information, and simple (weighted) concatenation is often performed (Schindler and van Gool, 2008).

Finally, the influence of the camera point of view and the classifier (NN and SVM) is analysed. To that end, we use the IXMAS dataset which has five different views (0–4) of the same action, and try some different choices as the sets of views available for training and testing. When the sets of training and test views have some common view, the examples of the shared views are randomly split into training an test sets in an 80%-20% ratio, and the average accuracy over 10 runs is computed. A linear SVM (Chang and Lin, 2011) is used and the regularization factor C is chosen from the set $\{10^e : e \in \{-3,-2,\ldots,4\}\}$.

Table 3: LOO accuracy (%) with NN using several frame and RM descriptors.

Frame descriptor ▽	RM descriptor ▽	Dataset		
		Weizmann	UIUC1	IXMAS
SH	HoL	37.8	76.1	34.7
	PaD	40.0	**83.8**	**66.9**
OF	HoL	50.0	69.9	45.6
	PaD	27.7	71.6	64.9
FFW	HoL	**55.6**	73.9	34.3
	PaD	46.7	63.4	59.9

Table 4: LOO and LOAO accuracies (%) with NN for different feature combinations.

Dataset ▷	Weizmann				UIUC1			
RM descriptor ▷	HoL		PaD		HoL		PaD	
Feature combination ▽	LOO	LOAO	LOO	LOAO	LOO	LOAO	LOO	LOAO
$\Phi_1 = \{\langle \mathsf{sh} \rangle, \langle \mathsf{of} \rangle\}$	48.9	53.3	36.7	40.0	80.3	60.0	85.3	63.2
$\Phi_2 = \{\langle \mathsf{sh}, \mathsf{of} \rangle\}$	57.8	61.1	46.7	48.9	80.1	60.2	85.5	64.7
$\Phi_3 = \{\langle \mathsf{ffw}, \mathsf{sh}, \mathsf{of} \rangle\}$	**63.3**	**66.7**	44.4	50.0	**85.2**	65.2	**87.2**	**66.7**
$\Phi_4 = \{\langle \mathsf{ffw} \rangle, \langle \mathsf{sh}, \mathsf{of} \rangle\}$	58.9	60.0	**53.3**	**56.7**	77.8	67.1	82.9	62.6
$\Phi_5 = \{\langle \mathsf{ffw} \rangle, \langle \mathsf{sh} \rangle, \langle \mathsf{of} \rangle\}$	58.9	58.9	43.3	47.8	79.7	**69.4**	84.7	65.6

To select the value of C, a validation procedure is followed using a 80%-20% split of the training set. To speed up the experiments, if a common C was observed to be consistently selected during some preliminary validation procedure, this value for C was fixed so that repetitive search is avoided.

Results (Table 5) show that that performance is generally better with SVM than with NN. This is particularly true when training and testing with sequences taken with the same camera (view). This seemingly counter-intuitive result can be due to the higher sensitiveness of the NN classifier to limited number of training examples. (When using the same view for training and testing, fewer examples are available for training). Besides the benefit of having more training examples, the fact of having higher performance when training and testing on disjoint sets of views suggests that the proposed RM descriptor exhibit some view invariance. It is worth noticing how the use of Φ_4 significantly outperforms Φ_3, most notably for the NN classifier. This indicates the positive effect of fusing the scores of classifiers based on different sets of features. Please, note that no late fusion is involved in Φ_3, since it consists of a single set of concatenated RM descriptors.

As a rough comparison with state-of-the-art similar approaches, results from (Junejo et al., 2011) are also provided in Table 5 when they use a combination of HOG and optic flow. While their results are usually better when using the same view for training and testing, our proposal performs similarly or better in other view combinations. Although the comparison is not performed under exactly the same conditions[2], the figures suggest that our proposal for RM descriptors and feature combinations is competitive, even with simpler descriptors and classifiers (e.g. they use a non-linear SVM with a χ^2 kernel).

4 CONCLUSIONS

A temporally holistic action representation based on recurrence matrices has been explored. Two recurrence matrix descriptors, and a general way of feature generation which combines early and late fusion strategies, have been proposed. The experiments reveal that the proposed descriptors offer competitive results despite being simpler than an existing one in the context of recurrence matrices. However, the performance depends on having enough training examples and/or using advanced classifiers. Further work

[2]For instance, (Junejo et al., 2011) reports not to have the same performer in the training and test sets at the same time. However, we did not consider this separation since the annotation of the performer was not found available in the feature dataset (Tran-Sorokin, 2008) that we use.

Table 5: Accuracy (%) with classifiers NN and SVM ($C = 10^4$), RM descriptor PaD, and feature combination Φ_3 and Φ_4 (see Table 4) on different sets of views (IXMAS dataset) for training and testing.

		Training Views : Testing Views							
Feat. Comb. ▽	Classifier ▽	0:0	1:1	2:2	3:3	4:4	2:3	1,2:3	All:All
Φ_3	NN	44.9	49.5	45.5	45.9	47.3	66.4	65.4	66.6
	SVM	65.8	63.9	64.5	65.1	61.1	68.9	80.6	78.4
Φ_4	NN	66.3	66.9	69.1	66.6	60.5	79.5	81.1	71.2
	SVM	70.5	72.3	73.0	75.4	65.8	68.9	77.0	75.3
SSM (HOG+OF) (Junejo et al., 2011)		77.0	77.3	75.8	71.2	68.8	68.5	N/A	74.6

is required to understand why the proposed system (descriptors, fusion strategy, or classifier) is somehow behind the state-of-the-art results, so that it can be made more discriminative, yet as simple as possible.

Combining features of different nature (such as shape, motion, and time-contextual information) generally improves the performance over individual subsets of these features. However, it is observed that which frame descriptors are chosen and how they are combined may significantly affect the performance in a data-dependent way. Consequently, devising an efficient procedure to select both, a proper subset of the descriptor parts, and a suitable fusion strategy, is among the most interesting research possibilities.

ACKNOWLEDGEMENTS

This work is partially supported by the Spanish research programme Consolider Ingenio-2010 CSD2007-00018, Fundació Caixa-Castelló Bancaixa (projects P1·1A2010-11 and P1·1B2010-27), and Generalitat Valenciana (PROMETEO/2010/028).

REFERENCES

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Comp. Surv.*, 43(3).

Ali, S., Basharat, A., and Shah, M. (2007). Chaotic invariants for human action recognition. In *ICCV*.

BenAbdelkader, C., Cutler, R., and Davis, L. S. (2004). Gait recognition using image self-similarity. *EURASIP J. on Applied Signal Processing*, 2004(4).

Brendel, W. and Todorovic, S. (2010). Activities as time series of human postures. In *ECCV*, pages 721–734.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Cutler, R. and Davis, L. S. (2000). Robust periodic motion and motion symmetry detection. In *CVPR*, pages 2615–2622.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.

Gaidon, A., Harchaoui, Z., and Schmid, C. (2011a). Actom sequence models for efficient action detection. In *CVPR*, pages 3201–3208.

Gaidon, A., Harchaoui, Z., and Schmid, C. (2011b). A time series kernel for action recognition. In *BMVC*.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *PAMI*, 29(12):2247–2253.

Junejo, I. N., Dexter, E., Laptev, I., and Pérez, P. (2011). View-independent action recognition from temporal self-similarities. *PAMI*, 33(1):172–185.

Lan, Z.-z., Bao, L., Yu, S.-I., Liu, W., and Hauptmann, A. G. (2012). Double fusion for multimedia event detection. In *Proc. of the 18th Intl. Conf. on Advances in Multimedia Modeling*, pages 173–185.

Lucena, M. J., de la Blanca, N. P., and Fuertes, J. M. (2012). Human action recognition based on aggregated local motion estimates. *Mach. Vis & Apps. (MVA)*, 23(1):135–150.

Marwan, N., Romano, M. C., Thiel, M., and Kurthss, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5–6):237–329.

Matikainen, P., Hebert, M., and Sukthankar, R. (2010). Representing pairwise spatial and temporal relations for action recognition. In *ECCV*, pages 508–521.

Niebles, J. C., Chen, C.-W., and Li, F.-F. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, pages 392–405.

Schindler, K. and van Gool, L. (2008). Action snippets: How many frames does human action recognition require? In *CVPR*.

Serra-Toro, C. and Traver, V. J. (2011). A new pedestrian detection descriptor based on the use of spatial recurrences. In *CAIP*, pages 97–104.

Tran, D. and Sorokin, A. (2008). Human activity recognition with metric learning. In *ECCV*, pages 548–561.

Tran-Sorokin (2008). Human activity recognition with metric learning. http://vision.cs.uiuc.edu/projects/activity.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Comp. Vis. & Image Underst. (CVIU)*, 104(2–3):249–257.