HAL
open science

Action Recognition using Exemplar-based Embedding

Daniel Weinland, Edmond Boyer

To cite this version:

Daniel Weinland, Edmond Boyer. Action Recognition using Exemplar-based Embedding. CVPR 2008 - IEEE Conference on Computer Vision and Pattern Recognition, Jun 2008, Anchorage, United States. pp.1-7, 10.1109/CVPR.2008.4587731 . inria-00590256

HAL Id: inria-00590256

https://inria.hal.science/inria-00590256

Submitted on 3 May 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Action Recognition using Exemplar-based Embedding

Daniel Weinland* Edmond Boyer
LJK - INRIA Rhône-Alpes, France
{weinland, eboyer}@inrialpes.fr

Abstract

In this paper, we address the problem of representing human actions using visual cues for the purpose of learning and recognition. Traditional approaches model actions as space-time representations which explicitly or implicitly encode the dynamics of an action through temporal dependencies. In contrast, we propose a new compact and efficient representation which does not account for such dependencies. Instead, motion sequences are represented with respect to a set of discriminative static key-pose exemplars and without modeling any temporal ordering. The interest is a time-invariant representation that drastically simplifies learning and recognition by removing time related information such as speed or length of an action. The proposed representation is equivalent to embedding actions into a space defined by distances to key-pose exemplars. We show how to build such embedding spaces of low dimension by identifying a vocabulary of highly discriminative exemplars using a forward selection. To test our representation, we have used a publicly available dataset which demonstrates that our method can precisely recognize actions, even with cluttered and non-segmented sequences.

1. Introduction

Action recognition is of central importance in computer vision with many applications in visual surveillance, human computer interaction and entertainment, among others. A challenging issue in this field originates from the diversity of information which describes an action. This includes purely visual cues, e.g. shape and appearance, as well as dynamic cues, e.g. space-time trajectories and motion fields. Such diversity raises the question of the relative importance of these sources and also to what degree they compensate for each other.

In a seminal work, Johansson [15] demonstrated through psychoanalytical experiments that humans can recognize actions merely from the motion of a few light points attached to the human body. Following this idea, several works, *e.g.* [1, 11], attempted to recognize actions using trajectories of markers with specific locations on the human body. While successful in constrained environments, these approaches do not however extend to general scenarios.

Besides, static visual information give also very strong cues on activities. In particular, humans are able to recognize many actions from a single image (see for instance Figure 1). Consequently a significant effort has been put in representations based on visual cues. Two main directions have been followed. *Implicit representations* simultaneously model in space and time with space-time volumes, e.g. [3, 25], or by using space-time features, e.g. [6, 18, 19]. *Explicit representations* equip traditional temporal models, such as hidden Markov models (HMMs), with powerful image matching abilities based on exemplar representations, e.g. [21, 7, 8, 24].

In this work we take a different strategy and represent actions using static visual information without temporal dependencies. Our results show that such representations can effectively model complex actions and yield recognition rates that equal or exceed those of the current state-of-the-art approaches, with the virtues of simplicity and efficiency.

Our approach builds on recent works on example-based embedding methods [2, 12]. In these approaches complex distances between signals are approximated in a Euclidean embedding space that is spanned by a set of distances to exemplar measures. Our representations is grounded on such embedding, focusing only on the visual components of an action. The main contribution is a time-invariant representation that does not require a time warping step and is insensitive to variations in speed and length of an action. To the best of our knowledge, no previous work has attempted to use such an embedding based representation to model actions.

In the paper, we will show how to select exemplars for such a representation using a forward feature selection technique [16]. In particular, we will demonstrate how complex actions can be described in terms of a small but highly discriminative exemplar sets. Experiments on the well known

*D. Weinland is supported by a grant from the European Community under the EST Marie-Curie Project Visitor.

Figure 1. Sample images from the *Weizmann*-dataset [3]. A human observer can easily identify many, if not all, actions from a single image. The interested reader may recognize the following actions: *bend*, *jumping-jack*, *jump-in-place*, *jump-forward*, *run*, *gallop-sideways*, *walk*, *wave one hand*, *wave two hands*, and *jump-forward-one-leg*. Note, that the displayed images have been automatically identified by our method as discriminative exemplars.

Weizmann-dataset [3] confirm that action recognition can be achieved without considering temporal dependencies.

Another important feature of our approach is that it can be used with advanced image matching techniques, such as the Chamfer distance [10], for visual measurements. In contrast to the classical use of dimensional reduction with silhouette representations, *e.g*. [23], such a method can be used in scenarios where no background subtraction is available. In a second experiment we will demonstrate, that even on cluttered non-segmented sequences, our method has precise recognition results.

The paper is organized as follows: In Section 2 we review related work. In Section 3 we present the embedding representation. In Section 4 we show how to compute a small but discriminative exemplar set. In Section 5 we evaluate our approach with a publicly available dataset before concluding and discussing issues in Section 6.

2. Related Work

Actions can been recognized using the occurrences of *key-frames*. In the work of Carlsson and Sullivan [4], class representative silhouettes are matched against video frames to recognize forehand and backhand strokes in tennis recordings. In a similar way, our approach uses a set of representative silhouette like models, *i.e*. *the exemplars*, but does not assume a deterministic framework as in [4], where exemplars are exclusively linked to classes, and decisions are based on single frame detections.

Other exemplar based approaches, e.g. [7, 8, 21, 24], learn HMMs with observation probabilities based on matching distances to exemplars. In all these models, dynamics are explicitly modeled through Markovian transitions over discrete state variables, whereas distances are mapped onto probabilities, which can involve additional difficulties [17].

Dedeoglu *et al*. [5] propose a real-time system for action recognition based on key-poses and histograms. Histograms introduce some degree of temporal invariance, although temporal order remains partially constrained with such representation. Moreover, the conversion of exemplar distances into normalized distributions can cause additional loss.

Exemplar based embedding methods have already been proposed, e.g. [2, 12]. In [2] Athitsos and Sclaroff present an approach for hand pose estimation based on Lipschitz embeddings. Guo *et al*. [12] use an exemplar-base embedding approach to match images of cars over different viewpoints. However no attempts has been made to apply such exemplar-based embedding approaches to action recognition.

Wang and Suter [23] use kernel-PCA to derive a low dimensional representation of silhouettes, and factorial conditional random fields to model dynamics. Having similar results in evaluation than our method, such an approach is computationally expensive, and moreover only practical in background subtracted scenes.

Interestingly, the *S3-C3 stage* of the biological motivated system by Jhuang *et al*. [14] shares as well some similarities with our embedding representation. However, these two representation are derived in a very different context.

3. Action Modeling

Our approach proceeds as illustrated in Figure 2. An action sequence is matched against a set of n exemplars. For each exemplar the minimum matching distance to any of the frames of the sequence is determined, and the resulting set of distances forms a vector D^* in the embedding space $\mathbb{R}^n$. The intuition we follow is that similar sequences will yield proximities to discriminative exemplars which are similar. Hence their point representation in $\mathbb{R}^n$ should be close. We thus model actions in $\mathbb{R}^n$ where both learning and recognition are performed. This is detailed in the following sections.

3.1. Exemplar-based Embedding

Our aim is to classify an action sequence $Y = y_1, \ldots, y_t$ over time with respect to the occurrence of known representative exemplars $X = \{x_1, \ldots, x_n\}$, e.g. silhouettes. The exemplar selection is presented in a further section (see Section 4) and we assume here that they are given.

We start by computing for each exemplar x_i the minimum distance to frames in the sequence:

$$d_i^*(Y) = \min_j d(x_i, y_j), \tag{1}$$

where d is a distance function between the primitives considered, as described in Section 3.3.

At this stage, distances could be thresholded and converted into binary detections, in the sense of a *key-frame* classifier [4]. This requires however thresholds to be chosen and furthermore does not allow to model uncertainties.

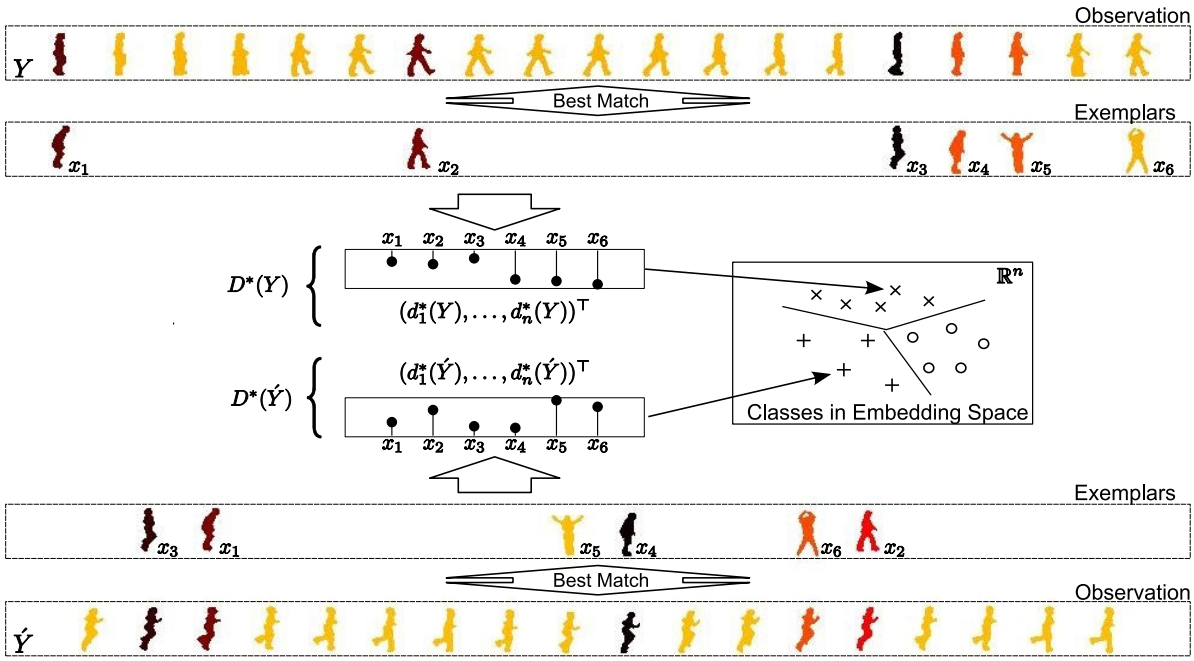

Figure 2. Overview of the embedding method: Two action sequences Y (*walk*) and $\acute{Y}$ (*jump forward on one leg*) are matched against a set of silhouette exemplars x_i. For each exemplar the best matching frame in the sequence is identified (exemplar displayed on top of the corresponding frame; light colors correspond to high matching distances; dark colors to low matching distances). The resulting matching distances d_i^* form vector D^*, which is interpreted as an embedding of the sequences into a low dimensional space $\mathbb{R}^n$. The final classifier is learned over $\mathbb{R}^n$, where each point represents a complete sequence.

Probabilistic exemplar-based approaches [21] do model such uncertainties by converting distances into probabilities, but as mentioned earlier, at the price of complex computations for normalization constants. We instead simply work on the vectors that result from concatenating all the minimum distances

$$D^*(Y) = (d_1^*(Y), \ldots, d_n^*(Y))^\top \in \mathbb{R}^n, \tag{2}$$

without any probabilistic treatment. Note that our representation is similar in principle to the embedding described in [2, 12] in a static context. We extend it to temporal sequences.

3.2. Classifier

In the embedding space $\mathbb{R}^n$, classification of time sequences reduces to a simple operation which is to label the vectors $D^*(Y)$. A major advantage over traditional approaches is that such vectors encode complete sequences without the need for time normalizations or alignments. These vectors are points in $\mathbb{R}^n$ that are labelled using a standard Bayes classifier. Each class $c \in 1...C$ is represented through a single Gaussian distribution $p(D^*|c) = \mathcal{N}(D^*|\mu_c, \Sigma_c)$, which we found adequate in experiments to model all important dependencies between exemplars. Assignments are determined through maximum a posteriori estimations:

$$g(D^*) = \arg\max_c p(D^*|c)p(c), \tag{3}$$

with $p(c)$ being the prior of class c that, without loss of generality, is assumed to be uniform.

Note that when estimating covariance Σ, and depending on the dimension n, it is often the case that insufficient training data is available for Σ, and consequently the estimation may be non-invertible. We hence work with a regularized covariance of the form $\hat{\Sigma} = \Sigma + \epsilon I$, with I being the identity matrix and ϵ a small value.

3.3. Image Representation and Distance Functions

Actions are represented as vectors of distances from exemplars to the frames in the action's sequence. Such distances could be of several types, depending on the available information in the images, e.g. silhouettes or edges. In the following, we assume that silhouettes are available for the exemplars, which is a reasonable assumption in the learning phase, and we consider two situations for recognition. First, silhouettes, obtained for instance with background subtractions, are available; Second only edges can be considered.

Silhouette-to-Silhouette Matching In this scenario we assume that background subtracted sequences are available.

Consequently, x and y are both represented through silhouettes. While difficult to obtain in many practical contexts, silhouettes, when available, provide rich and strong cues. Consequently they can be matched with a standard distance function and we choose the squared Euclidean distance $d(x, y) = |x - y|^2$, which is computed between the vector representations of the binary silhouette images. Hence, the distance is simply the number of pixels with different values in both images.

Silhouette-to-Edge Matching In a more realistic scenario, background subtraction will not be possible due to moving or changing background as well as changing light, among other reasons. In that case, more advanced distances dealing with imperfect image segmentations must be considered. In our experiments, we use such a scenario where edge observations y, instead of silhouettes, are taken into account. In such observations, edges are usually spurious or missing. As mentioned earlier we assume that exemplars are represented through edge templates, computed using background subtraction in a learning phase. The distance we consider is then the Chamfer distance [10],which measures the closest distance for each edge point on the observation x to any edge point in the exemplar y,

$$d(x, y) = \frac{1}{|x|} \sum_{f \in x} \mathsf{d}_y(f), \tag{4}$$

where $|x|$ is the number of edge points in x and $\mathsf{d}_y(f)$ is the distance between edge f and the closest edge-point in y. An efficient way to compute the Chamfer distance is by correlating the distance transformed observation with the exemplar silhouette.

4. Key-Pose Selection

In the previous section, we assume that the exemplars, a set of discriminative primitives, are known. We explain in this section how to obtain them using a forward feature selection. In a classical way, such selection has to deal with two conflicting objectives. First, the set of exemplars must be small to avoid learning and classification in high dimensions (*curse of dimensionality*) and to allow for fast computations. Second, the set must contain enough elements to account for variations within and between classes. We will use the wrapper technique for feature selection introduced in [16], but other possibilities will be discussed in Section 4.2

Several criteria exist to measure and optimize the quality of a feature set (see *e.g.* [13]). The wrapper approach can be seen as a direct and straightforward solution this problem. The criterion optimized is the validation of the considered classifier, which is itself used as a black box by the wrapper while performing a greedy search over the feature space. There are different search strategies for the wrapper and we use a *forward selection*, which we recently successfully applied in a similar setting [24].

4.1. Forward Selection

Forward selection is a bottom-up search procedure that adds new exemplars to the final exemplar set one at a time until the final set is reached. Candidate exemplars are all frames in the training set, or a sub-sampled set of these frames. In each step of the selection, classifiers for each candidate exemplar set are learned and evaluated. Consequently, in the first iteration classifier for each single candidate exemplar are learned, the exemplar with the best evaluation performance is added to the final exemplar set, and the learning and evaluation step is repeated using pairs of exemplars (containing the already selected), triples, quadruples, *etc*. The algorithm is given below (see Algorithm 1).

Algorithm 1 Forward Selection

Input: training sequences $\mathcal{Y} = \{Y_1, \dots Y_m\}$, validation sequences $\hat{\mathcal{Y}} = \{Y_1, \dots Y_{\hat{m}}\}$

1. let candidate exemplar set $\mathcal{X} = \{y : y \in \mathcal{Y}\}$
2. let final exemplar set $X = \emptyset$
3. while size of X smaller than n
   (a) for each $y \in \mathcal{X}$
      i. set $X' \leftarrow \{y\} \cup X$
      ii. train classifier g with $\mathcal{Y}$ and keep validation performance on $\hat{\mathcal{Y}}$
   (b) set $X \leftarrow \{y^*\} \cup X$ where y^* corresponds to the best validation performance obtained in step 3(a). If multiple y^* with same performance exist, randomly pick one.
   (c) set $\mathcal{X} \leftarrow \mathcal{X} \setminus \{y^*\}$
4. return X

4.2. Selection Discussion

Many techniques have been used in the literature to select exemplars and vocabulary sets in related approaches. For instance, several methods sub-sample or cluster the space of exemplars, *e.g.* [2, 21]. While generally applicable in our context, such methods require nevertheless very large sets of exemplars in order to reach the performance of a smaller set that has been specifically selected with respect to an optimization criterion. Moreover, as we observed in [24], a clustering can miss important discriminative exemplars, *e.g.* clusters may discriminate body shapes instead of actions.

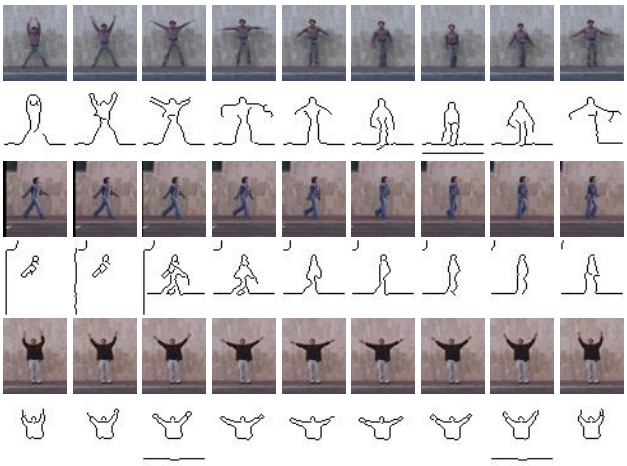

Figure 3. Sample sequences and corresponding edge images. (Top to bottom) *jumping-jack*, *walk*, *wave two hands*.

Another solution is to select features based on advanced classification techniques such as support vector machines [22] or Adaboost [9]. Unfortunately, support vector machines are mainly designed for binary classifications and, though extensions to multiple classes exist, they hardly extract a single feature set for all classes. On the other hand, Adaboost[9] can be extended to multiple classes and is known for its ability to search over large numbers of features. We experimented Adaboost using weak classifiers based on single exemplars and pairs of exemplars but performances were less consistent than with the forward selection.

Wrapper methods, such as the forward selection, are known to be particularly robust against over-fitting [13] but sometimes criticized for being slow due to the repetitive learning and evaluation cycles. In our case, we need approximately $n \times m$ learning and validation cycles to select n features out of a candidate set with size m. With a non-optimized implementation in MATLAB, selection of approximately 50 features out of a few hundreds will take around 5 minutes. This is a very reasonable computation time considering that this step is only required during the learning phase and that a compact exemplar set will benefit to all recognition phases.

5. Experiments

We have experimented our approach with the Weizmann-dataset [3] (see Figure 1 and 3) which has been recently used by several authors [1, 14, 18, 20, 23]. It contains 10 actions: *bend (bend)*, *jumping-jack (jack)*, *jump-in-place (pjump)*, *jump-forward (jump)*, *run (run)*, *gallop-sideways (side)*, *jump-forward-one-leg (skip)*, *walk (walk)*, *wave one hand (wave1)*, *wave two hands (wave2)*, performed by 9 actors. Silhouettes extracted from backgrounds and original image sequences are provided.

All recognition rates were computed with the leave-one-out cross-validation. Details are as follows. 8 out of the 9 actors in the database are used to train the classifier and select the exemplars, the 9th is used for the evaluation. This is repeated for all 9 actors and the rates are averaged. For the exemplar selection, we further need to divide the 8 training actors into training and validation sets. We do this as well with a leave-one-out cross-validation, using 7 training actors and the 8th as the validation set, then iterating over all possibilities. Exemplars are constantly selected from all 8 actors, but never from the 9th that is used for the evaluation. Also note that due to the small size of the training set, the validation rate can easily reach 100% if too many exemplars are considered. In this case, we randomly remove exemplars during the validation step, to reduce the validation rate and to allow new exemplars to be added. For testing we nevertheless use all selected exemplars.

5.1. Evaluation on Segmented Sequences

In these experiments, the background-subtracted silhouettes which are provided with the Weizmann-dataset were used to evaluate our method. For the exemplar selection, we first uniformly subsample the sequences by a factor $1/20$ and perform the selection on the remaining set of approximately 300 candidate frames. When we use all the 300 frames as exemplars, the recognition rate of our method is 100%.

To reduce the number of exemplars we search via forward selection over this set. In Figure 4, we show a sample exemplar set as returned from the selection method. Figure 4(a) shows the average validation rates per action, which were computed on the training set during the selection. Note that even though the overall validation rate reaches 100% for 15 exemplars, not all classes are explicitly represented through an exemplar, indicating that exemplars are shared between actions. The recognition rate on the test set and with respect to the number of exemplars is shown in Figure 4(b). Since the forward selection includes one random step, in the case where several exemplars present the same validation rate, we repeat the experiment 10 times with all actors, and average over the results. In Figure 4(c), we show recognition rates for the individual classes. Remark in particular the actions *jump-forward* and *jump-forward-one-leg* that are difficult to classify, because they are easily confused.

In summary, our approach can reach recognition rates up to 100% with approximately 120 exemplars. Moreover, with very small exemplar sets (e.g. around 20 exemplars), the average recognition rate on a dataset of 10 action and 9 actors is already higher than 90% and continuously increasing with additional exemplars (*e.g.* 97.7% for 50 exemplars). In comparison, the space-time volume ap-

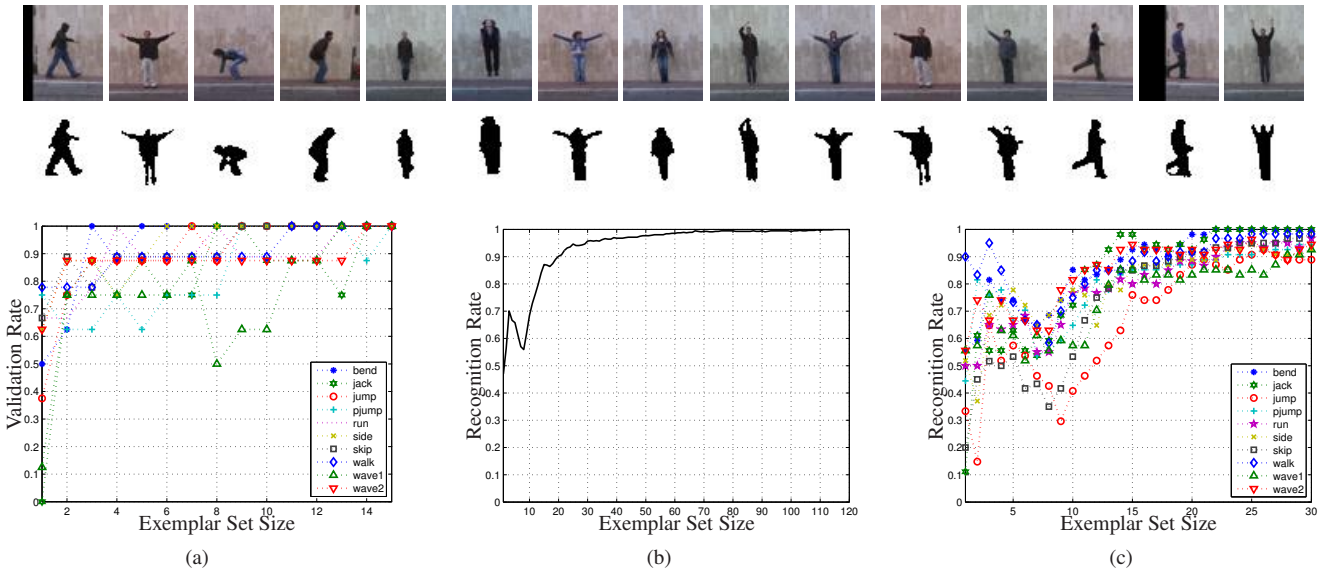

Figure 4. (Top) A set of exemplar silhouettes and their original images as returned by the forward selection (from left to right). (a) Corresponding validation rate per action during selection of exemplar set. (b) Recognition rates on background subtracted sequences vs. exemplar set size. (c) Recognition rates per action.

proach proposed by Blank *et al*. [3] had a recognition rate of 99.61%. Wang and Suter [23] report a recognition rate of 97.78% with an approach that uses kernel-PCA for dimensional reduction and factorial conditional random fields to model motion dynamics. The work of Ali *et al*. [1] uses a motion representation based on chaotic invariants and reports 92.6%. Note, however, that a precise comparison between the approaches is difficult, since experimental setups, *e.g*. number of actions and length of segments, slightly differ with each approach.

5.2. Evaluation on Cluttered Sequences

In this experiment, we used edge filtered sequences instead of background subtracted silhouettes. Edges are detected independently in each frame of the original sequences using a Canny edge detector. The resulting sequences contain a fair amount of clutter and missing edges, as can be seen in Figure 3. Exemplars are nevertheless represented through silhouettes since we assume that background subtraction is available during the learning phase though not during recognition. We also assume that the person centered region of interest in the image can be located.

For a uniformly sub-sampled exemplar set of size 300, our method presents a recognition rate of 93.6% in cross-validation on all 10 actions and 9 actors. Similarly to the previous experiment, we compute the recognition rate with respect to the number of selected exemplars. Figure 5(a) shows the average recognition rate, and Figure 5(b) the rate per action.

We observe that, after selection, a recognition rate of 93% can be achieved with 110 exemplars. Figure 5(c) shows the resulting confusion matrix in that case. As in the previous experiment, the two actions *jump-forward* and *jump-forward-one-leg* are difficult to classify, because they present many similarities. Another interesting observation is that, with only 2 exemplars, more than 50% of the actions are correctly classified.

In summary, our method shows very good results also on non-background subtracted sequences (up to 93.6% recognition rate). To our knowledge, methods that where tested on the Weizmann-dataset without using backgound subtraction are [14, 20, 18]. Jhuang *et al*. [14] report up to 98.8% recognition rate with their biologically motivated system. These results are however computed from only 9 actions and without the *jump-forward-one-leg* action which leads in our case to 4 false recognitions out of a total of 6. Scovanner *et al*. [20] mention 82.6% recognition rate using 3D SIFT descriptors and Niebles and Fei-Fei [18] 72.8% using spatial-temporal features. As in previous experiments, experimental setups are slightly different with each approach, *e.g*. [18, 20] additionally try to locate the person in the scene.

6. Conclusion and Discussion

We presented a new, compact, and highly efficient representation for action recognition. The representation is based on simple matching of exemplars to image sequences and does not account for dynamics. Based on exemplars, our representation supports advanced image matching distances and can be used with cluttered non-segmented sequences.

The experiments on sequences with and without back-

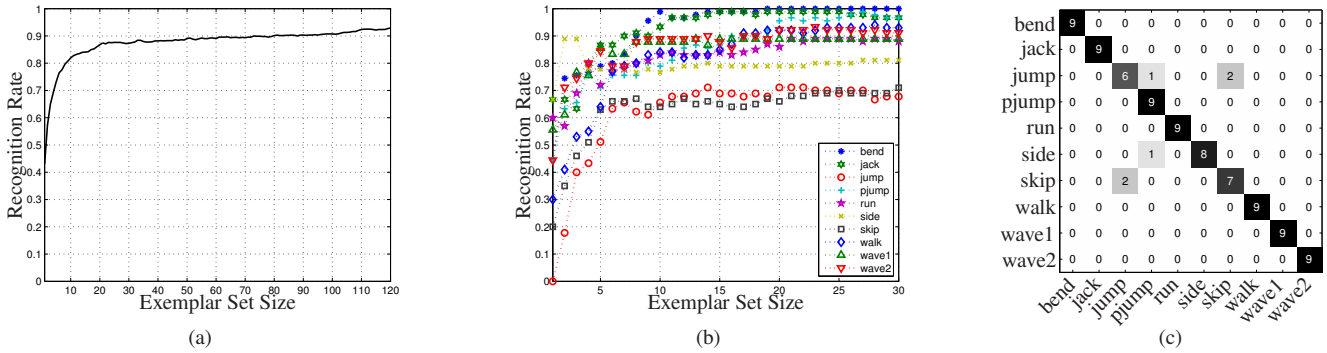

Figure 5. (a) Recognition rates vs. exemplar set size. Recognition is performed on non-background subtracted, edge filtered, sequences. (b) Recognition rates per action. (c) Confusion matrix using 110 exemplars

ground subtraction demonstrated that many actions can be recognized without taking dynamics into account. This was especially true on the publicly available Weizmann dataset, where our method has recognition rates which equal or exceed those of state-of-the-art approaches. To our opinion, this is an important result. However, it should be noticed that not all actions can be discriminated without dynamics. A typical example is an action and its reversal, *e.g.* *sit-down* and *get-up*. Without taking temporal ordering into account, it will be very difficult to discriminate them. To recognize such actions, a modeling of dynamics is required, either coupled with the descriptor or on a higher level. Nevertheless, note also that, as demonstrated with many of the datasets currently used in the field that do not include such ambiguous actions, many recognition applications do not need to discriminate such particular cases. On the other hand, we think that approaches could be experimented with more realistic scenes, to better evaluate their limitations.

We are currently working on temporal segmentation of sequences with our representation. Further, we are investigating exemplar selections in scenarios where no background subtraction is available during learning as well. We are also experimenting with other image-based matching distances.

References

[1] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *ICCV*, 2007. 1, 5, 6

[2] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *CVPR*, 2003. 1, 2, 3, 4

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005. 1, 2, 5, 6

[4] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. In *Workshop on Models versus Exemplars in Computer Vision*, 2001. 2

[5] Y. Dedeoglu, B. Toreyin, U. Gudukbay, and A. Cetin. Silhouette-Based Method for Object Classification and Human Action Recognition in Video. In *HCI*, 2006. 2

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, 2005. 1

[7] A. M. Elgammal, V. D. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition . In *CVPR*, 2003. 1, 2

[8] A. Fathi and G. Mori. Human pose estimation using motion exemplars. In *ICCV*, 2007. 1, 2

[9] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT* , 1995. 5

[10] D. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In *ICCV*, 1999. 2, 4

[11] N. H. Goddard. *The Perception of Articulated Motion: Recognizing Moving Light Displays*. PhD thesis, 1992. 1

[12] Y. Guo, Y. Shan, H. Sawhney, and R. Kumar. Peet: Prototype embedding and embedding transition for matching vehicles over disparate viewpoints. In *CVPR*, 2007. 1, 2, 3

[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3, 2003. 4, 5

[14] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action. In *ICCV*, 2007. 2, 5, 6

[15] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 1414, 1973. 1

[16] R. Kohavi and G. H. John. Wrappers for feature subset selection. *AI*, 97, 1997. 1, 4

[17] T. Minka. Exemplar-based likelihoods using the pdf projection theorem. Technical report, 2004. 2

[18] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, 2007. 1, 5, 6

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004. 1

[20] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Multimedia*, 2007. 5, 6

[21] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, 2001. 1, 2, 3, 4

[22] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998. 5

[23] L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *CVPR*, 2007. 2, 5, 6

[24] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, 2007. 1, 2, 4

[25] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *CVPR*, 2005. 1