SPARSE REPRESENTATIONS FOR HAND GESTURE RECOGNITION

Stergios Poularakis[⋆]*, Grigorios Tsagkatakis*[†]*, Panagiotis Tsakalides*[† ‡] *and Ioannis Katsavounidis*[⋆]

[⋆] Department of Computer and Communication Engineering, University of Thessaly, Greece
[†] Foundation for Research and Technology - Hellas (FORTH-ICS), Crete, Greece
[‡] Department of Computer Science, University of Crete, Greece

ABSTRACT

Dynamic recognition of gestures from video sequences is a challenging task due to the high variability in the characteristics of each gesture with respect to different individuals. In this work, we propose a novel representation of gestures as linear combinations of the elements of an overcomplete dictionary, based on the emerging theory of sparse representations. We evaluate our approach on a publicly available gesture dataset of Palm Grafti Digits and compare it with other state-of-the-art methods, such as Hidden Markov Models, Dynamic Time Warping and the recently proposed distance metric termed Move-Split-Merge. Our experimental results suggest that the proposed recognition scheme offers high recognition accuracy in isolated gesture recognition and a satisfying robustness to noisy data, thus indicating that sparse representations can be successfully applied in the field of gesture recognition.

Index Terms— gesture recognition, sparse representations, compressive sensing

1. INTRODUCTION

Gesture recognition is an active research area in the Computer Vision community, with applications in Human-Computer Interaction (HCI), sign language recognition and remote control of electronic devices. The release of mass consumer applications and devices, including gesture-controlled interactive TV systems (iDTV) and Microsoft's Kinect ™system, has fuelled the interest in gesture recognition technology.

Currently, there is a variety of models for representing gestures, with quite promising results [1]. Some of the most successfull include Hidden Markov Models (HMMs) [2] [3], Dynamic Time Warping (DTW) [4] [5], Conditional Random Fields (CRFs) [6] [7] and Dynamic Bayesian Networks (DBNs) [8]. Recently, some new approaches based on the emerging field of sparse representations and Compressive Sensing have also been proposed [10].

This research was partly funded by the PEOPLE-IAPP AVID-MODE and CS-ORION grants, within the 7th European Community Framework Program.

In this work, we focus on dynamic hand gestures, *i.e.* gestures where information lies in the trajectory of the hand palm. We represent gestures as sequences of hand coordinates on the image plane and formulate the recognition process as an instance of the sparse representation-based Classifier (SRC) [11]. Based on this formulation, class information is extracted by matching the sequence in an overcomplete dictionary of training examples. Experimental results on a publicly available dataset [4] suggest that the proposed system is able to achieve high recognition accuracy under challenging conditions.

In short, the novelty of our work is two-fold. First, we propose a gesture recognition algorithm that can achieve state-of-the-art performance in an efficient and robust way. This goal is achieved by leveraging the classification power of sparse representations for time series analysis. Second, the proposed scheme focuses on the combination of a strong classifier with a lightweight feature extraction, as opposed to more elaborate and time consuming feature extraction methods, thus offering greater design flexibility and more reasonable computational requirements.

The rest of this work is organized as follows: In Section 2 we review some related approaches for action recognition and sparse representations. In Section 3 we briefly present HMMs, DTW and MSM [12], which are used for comparison with our approach. In Section 4 we present our approach in detail and in Section 5 we present the experimental results. Finally, Section 6 concludes this work and addresses our plans for future work.

2. RELATED WORK

Our work is closely related to approaches that utilize the theory of sparse representations and overcomplete dictionaries for recognition purposes. According to this modelling paradigm, that was first introduced by Wright *et al.* [11] in the context of face recognition, a test signal can be represented as a linear combination of a few training examples from the same class. The notion of sparsity in the representation is mathematically formulated as a constrained least squares problem where the solution can be found by solving

an ℓ_0– or ℓ_1–minimization problem. The sparse representation framework was subsequently applied to other problems in addition to face recognition including generic human action recognition [13] [14] or unusual event detection in video sequences [15].

The work of Akl *et al.* [10] is the closest to our work. In that work, sparse representation of the hand coordinates in 3-D was used and an overcomplete dictionary was built, performing gesture recognition in a way similar to [11]. In our work, we follow a similar approach, but we use a different signal representation based on signal resampling. The proposed resampling mechanism addresses one of the main challenges in gesture recognition, the variability of duration between different gestures and different users, as well as different video frame rates. Furthermore, we restrict ourselves to the (x, y) hand coordinates, as acquired by a typical 2D vision-based hand detection mechanism, instead of (x, y, z), as acquired by a 3-axis accelerometer.

3. A BRIEF OVERVIEW OF HMMS, DTW AND MSM

In this work we address the issues of dynamic hand gesture recognition by posing the problem as time series classification, since information is encoded as a sequence of hand coordinates in time. One of the key challenges in classifying time series lies in the variability of each recorded signal with respect to its duration. In the context of gesture recognition, this phenomenon is illustrated by the different length of time that each individual takes to the perform a specific sequence. This variability is evident even when the same individual performs an action repeatedly. We discuss the state-of-the-art methods in time series classification next.

3.1. Hidden Markov Models

A Hidden Markov Model (HMM) [16] is a directed probabilistic graphical model that expresses the values of a time series by a set of hidden (unobservable) states of the learnt model. The three fundamental problems of HMMs are Evaluation, Decoding and Training and are commonly solved by the Forward [16], Viterbi [17] and Baum-Welch [18] algorithms, respectively. The key parameter of a HMM is the number of states that will be used for each model. Recognition is based on identifying the maximum likelihood path of states that generated the sequence of observations under testing (e.g. a gesture) from all available classes. For our experiments we used a Left-Right (Bakis) HMM and chose the best number of states for each class, after thorough cross-validation.

3.2. Dynamic Time Warping

Dynamic Time Warping (DTW) [19] utilizes Dynamic Programming in order to recursively compute the alignment and the corresponding distance between two time series. Since DTW is an exemplar, *i.e.* non–parametric, method, it typically requires little or even no training, in contrast to HMMs, that require extensive training, but is much slower during the recognition step, since all training examples have to be checked. For this reason, a *model version* of DTW is commonly used in recent approaches, such as [4].

3.3. Move-Split-Merge metric

The Move-Split-Merge (MSM) metric is proposed in [12] by Stefan *et al.*. It computes the distance between two time series, a and b, by using three fundamental operations, namely Move, Split and Merge, in order to transform a to b. MSM's advantage compared to DTW is metricity, which allows MSM to be further combined with a number of methods for indexing, clustering and visualization, designed for metric spaces. MSM's free parameter is the cost c of every Split and Merge operation. For our experiments we chose $c = 100$ since it resulted in the highest performance out of the values $\{0.01, 0.1, 1, 10, 100\}$.

4. PROPOSED GESTURE RECOGNITION TECHNIQUE

4.1. Gesture representation

For each of the training video segments, we assume a reliable hand detection preprocessing step, which extracts the (x, y) hand coordinates at each frame $t = 1, \ldots, T$, resulting in two 1D signals, $x = [x^{(1)}, \ldots, x^{(T)}]$ and $y = [y^{(1)}, \ldots, y^{(T)}]$. In order to achieve translation invariance (and some robustness to noise), we subtract the mean values $(\bar{x}, \bar{y})$ from the original observations, thus forming the signal vectors $x_{inv} = [x^{(1)} - \bar{x}, \ldots, x^{(T)} - \bar{x}]$ and $y_{inv} = [y^{(1)} - \bar{y}, \ldots, y^{(T)} - \bar{y}]$.

Since different gestures –or even different instances of the same gesture– generally have different duration (as measured in frames), we apply resampling through linear interpolation of (x_{inv}, y_{inv}) in order to obtain signal vectors $(\dot{x}, \dot{y})$ of fixed length N, where N is chosen as a parameter. Then we form a single 1D signal v of length $2N$ by interleaving $(\dot{x}, \dot{y})$, such that $v = [\dot{x}^{(1)}, \dot{y}^{(1)}, \ldots, \dot{x}^{(N)}, \dot{y}^{(N)}]$.

Finally we arrange the training sequences as columns of a matrix D, such that $D = [v_{1,1}, \ldots, v_{1,m_1}, \ldots v_{C,1}, \ldots v_{C,m_C}]$, where C denotes the number of different gesture classes, $m_i, i = 1, \ldots C$ is the number of training examples for each class i and $v_{i,j}$ is the j-th example of the i-th class.

4.2. Gesture Recognition

For each new test video sequence, we apply the above process in order to obtain the corresponding 1D signal v_{test}. We then compute the sparse representation s of v_{test} over the dictionary D, by solving the following optimization problem:

$$\hat{s} = \arg\min \|s\|_0 \text{ subject to } v_{test} = Ds \quad (1)$$

The main assumption is that if v_{test} belongs to class $c \in \{1, \ldots, C\}$, it will approximately lie in the linear span of the training examples from the same class c [11]. However, in practice, non-zero coefficients may result for other classes too. Sparse Representation-based Classifier (SRC) [11] resolves the ambiguity by minimizing the reconstruction error (or *residual error*) $r_c(v_{test})$ based only on the coefficients $\tilde{s_c}$ of a certain class c, *i.e.* $\hat{c} = arg\min_{c \in C} r_c(v_{test})$, where $r_c(v_{test}) = \|v_{test} - D\tilde{s_c}\|_2$ and $\tilde{s_c}$ is a masked version of s with zeros everywhere except for the entries associated with class c.

Since the ℓ_0–minimization problem (1) is generally NP-complete, in practice, one can use greedy algorithms, such as the Orthogonal Matching Pursuit (OMP) [20], which solves the following optimization problem:

$$\hat{s} = \arg\min \|v_{test} - Ds\|_2 \text{ subject to } \|s\|_0 \leq T_0 \quad (2)$$

where s is generally referred as the *coefficient vector* and T_0 is a constraint to the number of the non-zero elements of s (we used $T_0 = 2$). An alternative approach is to solve a convex relaxation of (1), also known as *noisy Basis Pursuit* (BP) [21]:

$$\hat{s} = \arg\min \|s\|_1 \text{ subject to } \|v_{test} - Ds\|_2 \leq \epsilon \quad (3)$$

where ϵ is the accuracy tolerant allowed to the solution $\hat{s}$ (we used $\epsilon = 0.001$).

For our experiments we used and compared both algorithms and refer to them as SRC-BP and SRC-OMP. The "SparseLab" package [22] and the OMP-Box [23] provide efficient Matlab implementations for the BP and the OMP algorithms, respectively.

Fig. 1. Typical frames from the 10 Palm Grafti Digits dataset of [4]. User wears a green glove and gestures the 10 digits such that recognition is possible only by the trajectory of the moving hand.

5. EXPERIMENTAL RESULTS

5.1. Experimental Setup

For our experiments we used the 10 Palm Grafti Digits dataset of [4], corresponding to the 10 arabic numerals $0 - 9$. The dataset is split in three subsets, namely the "Training", "Easy" and "Hard" sets. For our purposes we chose to use the "Training" set, in which the users wear a green glove, in order to facilitate the process of data acquisition. The dataset contains data from 10 users and each user performs each gesture three time in each video clips, resulting in a total of 30 examples for each gesture. Some representative frames can be seen in Figure 1. In this work, the center of gravity of the gesturing hand is selected as a representative feature in each frame.

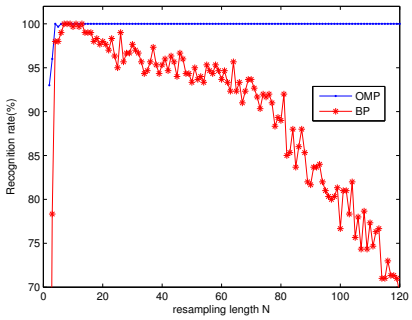

Fig. 2. Recognition accuracy of our approach as a function of the resampling parameter N.

In order to measure the recognition accuracy of both our approach and the three alternative approaches, we ran experiments in a leave-K-out cross-validation manner. In each round we considered K users (*i.e.* $M = 3K$ examples) as the training set and $10 - K$ users (*i.e.* $M = 3(10 - K)$ examples) as the validation set, while the final results correspond to the averaged accuracies.

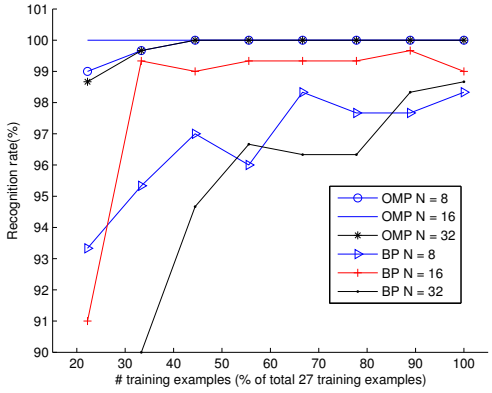

Fig. 3. Recognition accuracy of our approach as a function of the number of training examples for various values of the resampling parameter N.

5.2. Effect of resampling parameter N

The main parameter of our approach is the length N of the resampled sequence. As discussed in Section 4.1, this step is necessary in order to build the gesture dictionary. In our experiments we varied the value of this parameter in a wide range of values ($[1, 120]$). Our results for $K = 9$ users used as training set are shown in Figure 2, both for OMP and BP.

As it can be seen, our approach achieves high performance results even for really low values of N. More precisely, we observe that for $N \in [3, 40]$, BP achieves high accuracy, reaching 100%, while it starts diminishing slowly for $N > 40$. However, OMP presents an extremely stable behaviour and is not affected by N. In Figure 3 we confirm the same behaviour as we vary the number of training examples, M. OMP version is almost not affected at all by the choice of N and M, while BP is much less robust.

To our opinion, this is a very important result, since OMP is much more computationally efficient compared to BP. A further investigation of this issue is included in our goals for future work.

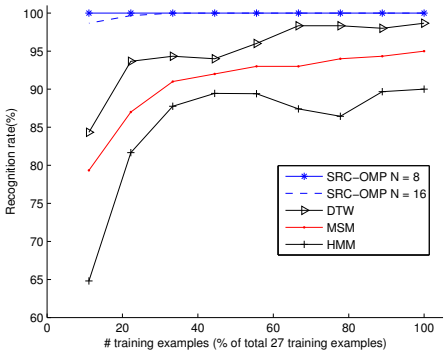

Fig. 4. Recognition accuracy of our approach and three other common methods as a function of the number of training examples (expressed as percentage of the total 27 examples).

5.3. Comparison with HMMs, DTW and MSM

We also ran experiments with the other three approaches, namely HMMs, DTW and MSM, in order to compare the results of our approach. Our main goal is to show that our approach produces good or even better results compared to the state–of–the–art. Comparison is based on the recognition accuracy as a function of the number of training examples. Our experimental results are shown in Figure 4. We should note that we present only the best results achieved by the other three methods (after thorough cross validation).

We observe that all four methods produce quite high results, while our approach outperforms the other three, achieving 100% accuracy even for 3 training examples (*i.e.* 1 user). We also observe that DTW produces the second best result, followed by MSM. HMMs show the worst performance among the methods presented, a result that is not surprising, given that this method typically requires a lot of training data. To our knowledge, the behaviour of DTW for few training examples has not been investigated so far.

5.4. Performance with noisy data

Since all of our current experiments are run in a dataset where the user wears a green glove, one can argue that the time series data we use is not fully representative of real–world data, obtained through real–time processing of video frames. Although our main goal is to demonstrate the recognition performance of our approach, we would like to provide a hint about the possible performance reduction due to noisy data. For this reason, we measured again the accuracy of all four methods, under the presence of additive white Gaussian noise $\mathcal{N}(0, \sigma^2)$ in our raw palm coordinate data (x, y). Figure 5 presents our results for $K = 9$ users.

Since HMMs performed very poorly, showing recognition rates as low as 30% for $SNR < 24dB$, we opted to compare the remaining methods for clarity. We observe that SRC-OMP, DTW and MSM present a very robust behaviour, with SRC-OMP providing the best results, even in the presence of very high noise (e.g. $SNR = 20dB$). Good performance of SR-based methods under high-noise conditions has also been confirmed in speech recognition too [9].

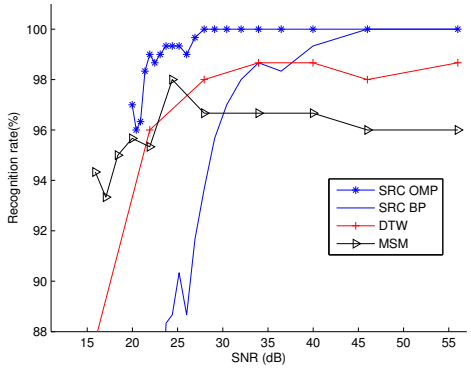

Fig. 5. Recognition accuracy of our approach in noisy data as a function of SNR.

6. CONCLUSION

In this work we proposed a new approach for gesture recognition based on sparse representations and an overcomplete gesture dictionary. Our experimental results show that our approach is meaningful and at least as accurate and robust as other commonly used (HMMs, DTW) or recently proposed (MSM) frameworks.

While our approach outperforms the other three approaches, it is worth saying that our primary goal is to demonstrate its competitive performance in gesture recognition tasks and not claim it to be a fundamentally better method. As it is also stated in [12], we should never forget that all these methods are valuable and the above results might be much different in a different dataset.

Our plans for future work include the evaluation of our approach in datasets without the limitations of the green glove, as well as the development of a framework for online gesture spotting instead of isolated gesture recognition, a problem in which HMMs and DTW present state-of-the-art results.

7. REFERENCES

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311–324, 2007.

[2] J. Ohya J. Yamato and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *IEEE Proc. Computer Vision and Pattern Recognition*, 1992, pp. 379–385.

[3] A. Al-Hamadi M. Elmezain and B. Michaelis, "Hand gesture spotting based on 3D dynamic features using hidden markov models," in *Signal Processing, Image Processing and Pattern Recognition*, 2009, pp. 9–16.

[4] Q. Yuan J. Alon, V. Athitsos and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 9, pp. 1685–1699, 2009.

[5] P. Radeva S. Escalera, A. Escudero and J. Vitrià, "Adaptive dynamic space time warping for real time sign language recognition," *4th CVC Workshop on Research and Development: New Trends and Challenges in Computer Vision*, pp. 155–160, 2009.

[6] A. Al-Hamadi M. Elmezain and B. Michaelis, "A robust method for hand gesture segmentation and recognition using forward spotting scheme in conditional random fields," in *IEEE Proc. 20th Int. Conf. Pattern Recognition (ICPR)*, 2010, pp. 3850–3853.

[7] S. Sclaroff H.D. Yang and S.W. Lee, "Sign language spotting with a threshold model based on conditional random fields," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 7, pp. 1264–1277, 2009.

[8] B.K. Sin H.I. Suk and S.W. Lee, "Hand gesture recognition based on dynamic bayesian network framework," *Pattern Recognition*, vol. 43, no. 9, pp. 3059–3072, 2010.

[9] J.F. Gemmeke T. Virtanen and A. Hurmalainen, "Exemplar based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[10] A. Akl and S. Valaee, "Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing," in *IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 2270–2273.

[11] A. Ganesh S.S. Sastry Y. J. Wright, A.Y. Yang and Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 31, no. 2, pp. 210–227, 2009.

[12] V. Athitsos A. Stefan and G. Das, "The move-split-merge metric for time series," *IEEE Trans. Knowledge and Data Engineering, (to appear)*, 2012.

[13] O.P. Concha R. Xu and M. Piccardi, "Compressive sensing of time series for human action recognition," *Int. Conf. Digital Image Computing: Techniques and Applications*, 2010.

[14] Y. Fu Y. Zhu, X. Zhao and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," *IEEE Int. Conf. Computer Vision (ACCV)*, pp. 660–671, 2010.

[15] Li Fei-Fei B. Zhao and E.P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3313–3320.

[16] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[17] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[18] G. Soules L.E. Baum, T. Petrie and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, pp. 164–171, 1970.

[19] J.B. Kruskall and M. Liberman, "The symmetric time warping algorithm: From continuous to discrete," *Time Warps*, pp. 125–162, 1983.

[20] R. Rezaiifar Y.C. Pati and P.S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," vol. 1, pp. 40–44, 1993.

[21] D.L. Donoho S.S. Chen and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[22] V. Stodden D. Donoho and Tsaig Y., "Sparselab 1.0," .

[23] Ron Rubinstein, "Ompbox v10," 2009.