Probability-based Dynamic Time Warping for Gesture Recognition on RGB-D data

Miguel Ángel Bautista[1,2], Antonio Hernández-Vela[1,2], Victor Ponce[1,2], Xavier Perez-Sala[2,3], Xavier Baró[2,4], Oriol Pujol[1,2], Cecilio Angulo[3], and Sergio Escalera[1,2]

[1] *Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain.*
[2] *Centre de Visió per Computador, Campus UAB, Edifici O, 08193 Bellaterra, Barcelona, Spain.*
[3] *CETpD-UPC, Universitat Politècnica de Catalunya Neàpolis, Rambla de l'Exposició, 59-69, 08800 Vilanova i la Geltrú, Spain*
[4] *EIMT, Universitat Oberta de Catalunya, Rambla del Poblenou 156, 08018, Barcelona, Spain.*

Abstract. *Dynamic Time Warping* (DTW) is commonly used in gesture recognition tasks in order to tackle the temporal length variability of gestures. In the DTW framework, a set of gesture patterns are compared one by one to a maybe infinite test sequence, and a query gesture category is recognized if a warping cost below a certain threshold is found within the test sequence. Nevertheless, either taking one single sample per gesture category or a set of isolated samples may not encode the variability of such gesture category. In this paper, a probability-based DTW for gesture recognition is proposed. Different samples of the same gesture pattern obtained from RGB-Depth data are used to build a Gaussian-based probabilistic model of the gesture. Finally, the cost of DTW has been adapted accordingly to the new model. The proposed approach is tested in a challenging scenario, showing better performance of the probability-based DTW in comparison to state-of-the-art approaches for gesture recognition on RGB-D data.

Keywords: Depth maps, Gesture Recognition, Dynamic Time Warping, Statistical Pattern Recognition.

1 Introduction

Nowadays, human gesture recognition is one of the most challenging tasks in computer vision. Current methodologies have shown preliminary results on very simple scenarios, but they are still far from human performance. Due to the large number of potential applications involving human gesture recognition in fields like surveillance [8], sign language recognition [10], or in clinical assistance [9] among others, there is a large and active research community devoted to deal with this problem.

The release of the Microsoft Kinect^{TM} sensor in late 2010 has allowed an easy and inexpensive access to synchronized range imaging with standard video data. This data combines both sources into what is commonly named RGB-D images (RGB plus Depth). This very welcomed addition by the computer vision community has reduced the burden of the first steps in many pipelines devoted to image or object segmentation and opened new questions such as how these data can be effectively fused and described. This depth information has been particularly exploited for human body segmentation and tracking. Shotton et. al [11] presented one of the greatest advances in the extraction of the human body pose using RGB-D, which is provided as part of the Kinect^{TM} human recognition framework. The extraction of body pose information opens the door to one of the most challenging problems nowadays, i.e. human gesture recognition. This fact has enabled researchers to apply new techniques to obtain more discriminative features. As a consequence, new methodologies on gesture recognition can improve their performance by using RGB-D data.

From a learning point of view, the problem of human gesture recognition is an example of sequential learning. The main problem in this scenario comes from the fact that data sequences may have different temporal duration and even be composed of intrinsically a different set of component elements. There are two main approaches for this problem: On the one hand, methods such as Hidden Markov Models (HMM) or Conditional Random Fields (CRF) are commonly used to tackle the problem from a probabilistic point of view [10], especially for classification purposes. Furthermore, methods based on key poses for gesture recognition have been proposed [6]. On the other hand, dynamic programming inspired algorithms can be used for both alignment and clustering of temporal series [5]. One of the most common dynamic programming methods used for gesture recognition is Dynamic Time Warping (DTW) [3,4].

However, the application of such methods to gesture recognition in complex scenarios becomes a hard task due to the high variability of environmental conditions. Common problems are: the wide range of human pose configurations, influence of background, continuity of human movements, spontaneity of humans actions, speed, appearance of unexpected objects, illumination changes, partial occlusions, or different points of view, just to mention a few. These effects can cause dramatic changes in the description of a certain gesture, generating a great intra-class variability. In this sense, since usual DTW is applied to compare a sequence and a single pattern, it fails when such variability is taken into account. We propose a probability-based extension of DTW method, able to perform an alignment between a sequence and a set of N pattern samples from the same gesture category. The variance caused by environmental factors is modelled using a Gaussian Mixture Model (GMM) [7]. Consequently, the distance metric used in the DTW framework is redefined in order to provide a probability-based measure. Results on a public and challenging computer vision dataset show a better performance of the proposed probability-based DTW in comparison to standard approaches.

The remaining of this paper is organized as follows: Section 2 presents the probability-based DTW method for gesture recognition, Section 4 presents the results and, finally, Section 5 concludes the paper.

2 Standard DTW for begin-end Gesture Recognition

In this section we first describe the original DTW and its common extension to detect a certain pattern sequence given a continuous and maybe infinite data stream. Then, we extend the DTW in order to align several patterns, taking into account the variance of the training sequence by means of a Gaussian mixture model.

2.1 Dynamic Time Warping

The original DTW algorithm was defined to match temporal distortions between two models, finding an alignment/warping path between the two time series $Q = \{q_1, .., q_n\}$ and $C = \{c_1, .., c_m\}$. In order to align these two sequences, a $M_{m \times n}$ matrix is designed, where the position (i, j) of the matrix contains the alignment cost between c_i and q_j. Then, a warping path of length τ is defined as a set of contiguous matrix elements, defining a mapping between C and Q: $W = \{w_1, .., w_\tau\}$, where w_i indexes a position in the cost matrix. This warping path is typically subjected to several constraints:

Boundary conditions: $w_1 = (1, 1)$ and $w_\tau = (m, n)$.

Continuity and monotonicity: Given $w_{\tau'-1} = (a', b')$, then $w_{\tau'} = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$, this condition forces the points in W to be monotonically spaced in time.

We are generally interested in the final warping path that, satisfying these conditions, minimizes the warping cost:

$$DTW(M) = \min_w \{M(w_\tau)\}, \tag{1}$$

where τ compensates the different lengths of the warping paths. This path can be found very efficiently using dynamic programming. The cost at a certain position $M(i, j)$ can be found as the composition of the Euclidean distance $d(i, j)$ between the feature vectors of the sequences c_i and q_j and the minimum cost of the adjacent elements of the cost matrix up to that point, i.e.:

$$M(i, j) = d(i, j) + \min\{M(i-1, j-1), M(i-1, j), M(i, j-1)\}. \tag{2}$$

Given the streaming nature of our problem, the input vector Q has no definite length and may contain several occurrences of the gesture pattern C. At that point the system considers that there is correspondence between the current block k in Q and a gesture if satisfying the following condition, $M(m, k) < \mu, k \in [1, .., \infty]$ for a given cost threshold μ.

This threshold value is estimated in advance using leave-one-out cross-validation strategy. This involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. At each iteration, we evaluate the similarity value between the candidate and the rest of the training set. Finally we choose the threshold value which is associated with the largest number of hits.

Once detected a possible end of pattern of gesture, the working path W can be found through backtracking of the minimum path from $M(m, k)$ to $M(0, z)$, being z the instant of time in Q where the gesture begins. Note that $d(i, j)$ is the cost function which measures the difference among our descriptors V_i and V_j.

An example of a begin-end gesture recognition together with the warping path estimation is shown in Figure 2.

3 Handling variance with Probability-based DTW

Consider a training set of N sequences $\{S_1, S_2, \ldots, S_N\}$, where each S_g represents a sample of the same gesture class. Then, each sequence S_g composed by a set of feature vectors at each time t, $S_g = \{s_1^g, \ldots, s_{L_g}^g\}$ for a certain gesture category, where L_g is the length in frames of sequence S_g. Let us assume that sequences are ordered according to their length, so that $L_{g-1} \leq L_g \leq L_{g+1}, \forall g \in [2, .., N-1]$, the median length sequence is $\bar{S} = S_{\lceil \frac{N}{2} \rceil}$. This sequence $\bar{S}$ is used as a reference, and the rest of sequences are aligned with it using the classical Dynamic Time Warping with Euclidean distance [3], in order to avoid the temporal deformations of different samples from the same gesture category. Therefore, after the alignment process, all sequences have length $L_{\lceil \frac{N}{2} \rceil}$. We define the set of warped sequences as $\tilde{S} = \{\tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_N\}$. Once all samples are aligned, the features vectors corresponding to each sequence element at a certain time t $\tilde{s}_t$ are modelled by means of an $G-$component Gaussian Mixture Model (GMM) $\lambda_t = \{\alpha_k, \mu_k, \Sigma_k\}, \quad k = 1, \ldots, G$, α is the mixing value and μ and Σ are the parameters of each of the G Gaussian models in the mixture. The underlying reason of choosing a GMM instead of a single Gaussian follows from the definition of the problem, where an arbitrarily large number of samples $\{S_1, S_2, \ldots, S_N\}$ is available. In this sense, in order to accurately model the feature vectors a GMM seems a more powerful way to model the variability than a single Gaussian. As a result, each one of the GMMs that model each component of a gesture pattern $\tilde{s}_t$ is defined as follows:

$$p(\tilde{s}_t) = \sum_{k=1}^{G} \alpha_k \cdot e^{-\frac{1}{2}(x-\mu_k)^T \cdot \Sigma_k^{-1} \cdot (x-\mu_k)}. \tag{3}$$

The resulting model is composed by the set of GMMs that model each one of the component elements among all warped sequences of a certain gesture class. An example of the process is shown in Figure 1.

3.1 Distance measures

In the classical DTW, a pattern and a sequence are aligned using a distance metric, such as the Euclidean distance. Since our gesture pattern is modelled by means of probabilistic models, if we want to use the principles of DTW, the distance needs to be redefined. In this paper we consider a soft-distance based on the probability of a point belonging to each one of the G components in the GMM, i.e., the posterior probability of x is obtained according to (3). In addition, since $\sum_{1}^{k} \alpha_k = 1$, we can compute the probability of an element $q \in Q$ belonging to the whole GMM λ as the following:

$$P(q, \lambda) = \sum_{k=1}^{M} \alpha_k \cdot P(q)_k, \tag{4}$$

$$P(q)_k = e^{-\frac{1}{2}(q-\mu_k)^T \cdot \Sigma_k^{-1} \cdot (q-\mu_k)}, \tag{5}$$

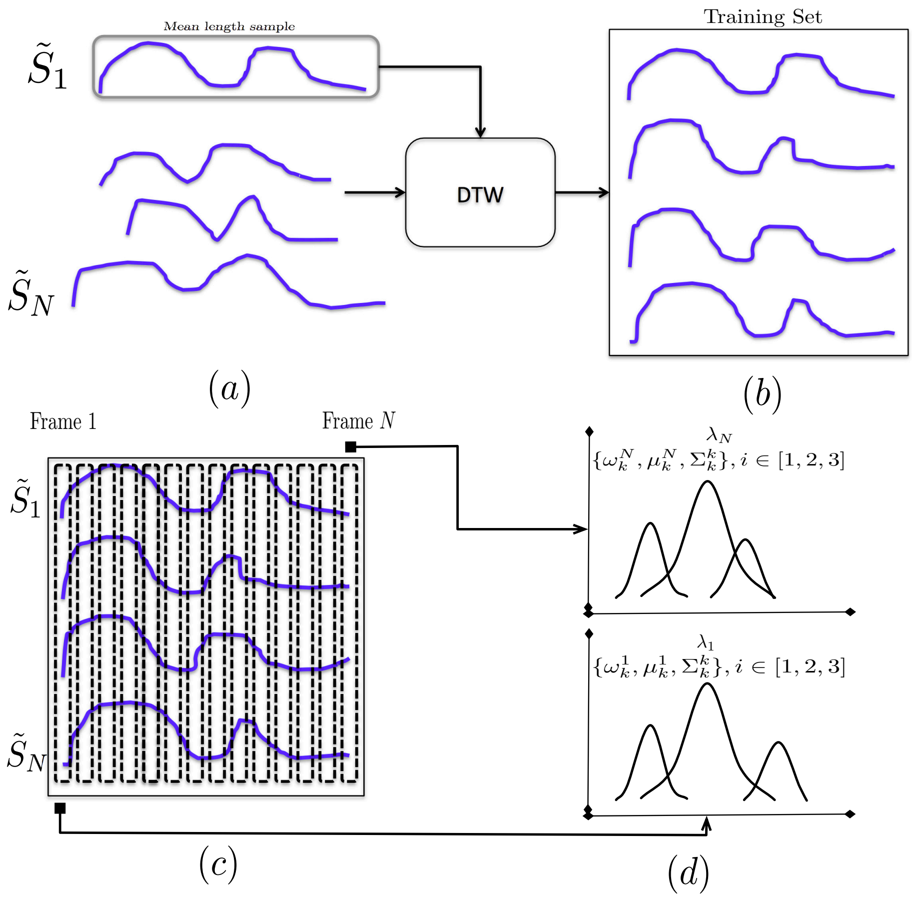

Fig. 1. (a) Different sample sequences of a certain gesture category and the mean length sample. (b) Alignment of all samples with the mean length sample by means of Euclidean DTW. (c) Warped sequences set $\tilde{S}$ from which each set of t-th elements among all sequences are modelled. (d) Gaussian Mixture Model learning with 3 components.

which is the sum of the weighted probability of each component. An additional step is required since the standard DTW algorithm is conceived for distances instead of similarity measures. In this sense, we use a soft-distance based measure of the probability, which is defined as:

$$D(q, \lambda) = e^{-P(q,\lambda)}. \tag{6}$$

In conclusion, possible temporal deformations of the gesture category are taken into account by aligning the set of N gesture sample sequences. In addition, modelling with a GMM each of the elements which compose the resulting warped sequences, we obtain a methodology for gesture detection that is able to deal with multiple deformations in data. The algorithm that summarizes the use of the probability-based DTW to detect start-end of gesture categories is shown in Table 1. Figure 4 illustrates the application of the algorithm in a toy problem.

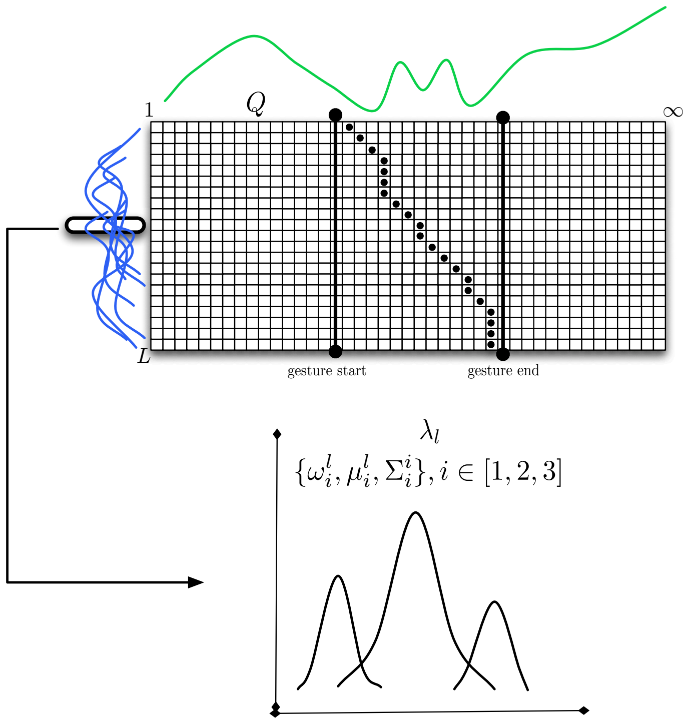

Fig. 2. Begin-end of gesture recognition of a gesture pattern in an infinite sequence Q using the probability-based DTW. Note that different samples of the same gesture category are modelled with a GMM and this model is used to provide a probability-based distance. In this sense, each cell of M will contain the accumulative D distance.

Input: A gesture model $C = \{c_1, .., c_m\}$ with corresponding GMM models $\lambda = \{\lambda_1, .., \lambda_m\}$, its similarity threshold value μ, and the testing sequence $Q = \{q_1, .., q_v\}$. Cost matrix $M_{m \times v}$ is defined, where $\mathcal{N}(x), x = (i, t)$ is the set of three upper-left neighbor locations of x in M.
Output: Working path W of the detected gesture, if any.
// Initialization
for $i = 1 : m$ **do**
    for $j = 1 : \infty$ **do**
        $M(i, j) = v$
end **end**
for $j = 1 : v$ **do**
    $M(0, j) = 0$
end
for $t = 0 : v$ **do**
    for $i = 1 : m$ **do**
        $x = (i, t)$
        $M(x) = D(q_t, \lambda_i) + \min_{x' \in \mathcal{N}(x)} M(x')$
    end
    if $M(m, t) < \epsilon$ **then**
        $W = \{\operatorname{argmin}_{x' \in \mathcal{N}(x)} M(x')\}$
        return
    end
end

Table 1. Probability-based DTW algorithm.

4 Experiments

In order to present the experiments, we discuss the data, methods and evaluation measurements.

4.1 Data

The data source used is the ChaLearn [2][1] data set provided from the CVPR2012 Workshop challenge on Gesture Recognition. The data set consists of 50,000 gestures each one portraying a single user in front of a fixed camera. The images are captured by the Kinect^{TM} device providing both RGB and depth images. The data used (a subset of the whole) are 20 development batches with a manually tagged gesture segmentation. Each batch includes 100 recorded gestures, grouped in sequences of 1 to 5 gestures performed by the same user. For each sequence the actor performs a resting gesture between each gesture of the gestures to classify. For this data set, we performed background subtraction based on depth maps, and we defined a 10×10 grid approach to extract HOG+HOF feature descriptors per cell, which are finally concatenated in a full image (posture) descriptor. In this data set we will test the recognition of the resting

[1] `http://gesture.chalearn.org/data/data-examples`

gesture pattern, using 100 samples of the pattern in a ten-fold validation procedure. An example of the ChaLearn dataset is shown in Figure 3.

(a) (b)

Fig. 3. Sample a) depth and b) RGB image for the ChaLearn database.

4.2 Methods and Evaluation

We compare the usual DTW and Hidden Markov Model (HMM) algorithms with our probability-based DTW approach using the proposed distance D shown in (6). The evaluation measurements are the accuracy of the recognition and the overlapping for the resting gesture (in percentage). We consider that a gesture is correctly detected if the overlapping in the resting gesture sub-sequence is grater than 60% (a standard overlapping value). The cost-threshold for all experiments was obtained by cross-validation on training data, using a 5-fold cross-validation, and the confidence interval was computed with a two-tailed t-test. Each GMM in the probability-based DTW was fit with $k = 4$ components, this value was obtained using a 2-fold cross-validation procedure on training data. For HMM, it was trained using the Baum-Welch algorithm, and 3 states were experimentally set for the resting gesture, using a vocabulary of 60 symbols computed using K-means over the training data features. Final recognition is performed with temporal sliding windows of different wide sizes, based on the training samples length variability.

Table 2 shows the results of HMM and the classical DTW algorithm, in comparison to our proposal on the ChaLearn dataset. We can see how the proposed probability-based DTW outperforms the usual DTW and HMM algorithms in both experiments. Moreover, confidence intervals of DTW and HMM do not intersect with the probability-based DTW in any case. From this results we can observe how performing dynamic programming increases the generalization capability of the HMM approach, as well as a model defined by a set of GMMs outperforms the classical DTW [3] on RGB-Depth data without increasing the computational complexity of the method.

Fig. 4. Examples of resting gesture detection on the Chalearn dataset using the probability-based DTW approach. The line below each pair of depth and RGB images represents the detection of a resting gesture.

	Overlap.	Acc.
Probability-based DTW	**39.08± 2.11**	**67.81±2.39**
Euclidean DTW	30.03±3.02	60.43± 3.21
HMM	28.51±4.32	53.28±5.19

Table 2. Overlapping and Accuracy results of different gesture recognition approaches.

5 Conclusions and Future Work

In this paper, we proposed a probability-based DTW for gesture recognition on RGB-D data, where the pattern model is learned from several samples of the same gesture category. Different sequences were used to build a Gaussian-based probabilistic model of the gesture whose possible deformations are implicitly encoded. In addition, a soft-distance based on the posterior probability of the GMM was defined. The novel approach has been successfully applied on a public RGB-D gestures dataset, being able to deal with multiple deformations in data, and showing performance improvements compared to the classical DTW and HMM approaches. In particular, the proposed method benefits from both the generalization capability from the probabilistic framework, when several observations of the training data are available, and the temporal warping capability from dynamic programming.

Future work lines include, between others, the inclusion of samples with different points of view of the same gesture class, the analysis of state-of-the-art one-class classifiers in order to obtain a performance improvement, and the definition of powerful descriptors to obtain gesture-discriminative features.

Acknowledgements

This work is partly supported by projects IMSERSO-Ministerio de Sanidad 2011 Ref. MEDIMINDER and RECERCAIXA 2011 Ref. REMEDI, and SUR-DEC of the Generalitat de Catalunya and FSE. The work of Antonio is supported by an FPU fellowship from the Ministerio de Educacion of Spain.

References

1. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43-72, 2005.
2. ChaLearn Gesture Dataset (CGD2011), ChaLearn, California, 2011. Copyright (c) ChaLearn - 2011
3. Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43-49, 1978.
4. M. Reyes, G. Dominguez, and S. Escalera. Feature weighting in dynamic time warping for gesture recognition in depth data. Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 1182–1188, 2011.
5. F. Zhou, F. D. la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 35, Issue 3, pp. 582-596, 2010.
6. Fengjun Lv; Nevatia, R., Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on , vol., no., pp.1,8, 17-2, 2007
7. M. Svensen and C. M. Bishop. Robust bayesian mixture modelling. *European Symposium on Artificial Neural Networks, Proceedings of*, 64:235–252, 2005.

8. A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti. Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking. *Signal Processing Magazine, IEEE*, 22(2):38–51, 2005.
9. A. Pentland. Socially aware computation and communication. *Computer*, 38:33–40, 2005.
10. Starner, T. and Pentland, A. Real-time American Sign Language recognition from video using hidden Markov models Computer Vision, 1995. Proceedings., International Symposium on, 265 -270
11. Shotton, Jamie and Sharp, Toby and Kipman, Alex and Fitzgibbon, Andrew and Finocchio, Mark and Blake, Andrew and Cook, Mat and Moore, Richard Real-time human pose recognition in parts from single depth images Communications of the ACM, volume 56, pp. 116–124, 2013.