Adaptive Dynamic Space Time Warping for Real Time Sign Language Recognition

Sergio Escalera, Alberto Escudero, Petia Radeva, and Jordi Vitrià
Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Spain UB
Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain

Abstract

The problem of automatic action recognition in uncontrolled environments becomes a hard because due to the high changes in action appearance because of illumination changes, frame resolution, occlusions, background moving objects, etc. In this paper, we propose a general framework for automatic action classification applied to the sign language recognition problem. The system is based on skin blob detection and tracking. Using a coordinate system estimated from a face detection procedure, the final sign is recognized by means of a new adaptive method of Dynamic Space Time Warping. Results over a Sign Language database show high performance improvement classifying more than 20 signs.
Keywords: Sign Language Recognition, Dynamic Space Time Warping.

1 Introduction

Automatic Action and gesture recognition is a challenging task is the fields of social signal processing, affective computing, communication or psychology, between others. In the case of sign language, recognition is a hard task because of the high changes of gestures in motion and appearance. Recent works try to deal with this problem by means of tracking blobs mainly corresponding to hands. Afterwards, temporal knowledge is used to perform sign classification.

Concerning the segmentation and feature extraction steps, several works use special clothes or cumbersome devices such as colored markers or gloves [1]. Common approaches for hand location base on skin detection, motion detection, edges, or background substraction [2, 3]. After localizing the regions of interest, motion information or local descriptors, such as SIFT [4] or HOG [5], are frequently used to describe the region content.

In our work, subjects can appear either with short-sleeved or long-sleeved clothes. In this domain, working with multiple region hypotheses is recommended. In order to work with multiple candidates, Sato and Kobayashi [6] extended the Viterbi algorithm in the Hidden Markov Model (HMM) accommodating multiple hypothesis at each query frame. Dynamic Space Time Warping (DSTW) [7] was defined as an extension of Dynamic Time Warping (DTW) [8] in order to deal with a fixed number of candidates by frame. Recently, Conditional Random Fields (CRF) have been also applied to sign language recognition in order to learn an adaptive threshold able to distinguish between vocabulary and non-sign patterns [9].

In this paper, we suppose that the number of candidates should vary based on the size of the segmented body region. This allows a problem-dependent adaptation that reduces time complexity while preserving (or even improving) the performance of the sign language recognition system. The scale and translation invariant approach, based on [10], defines a bottom-up procedure where skin

regions are segmented, described, and temporally recognized as signs of the vocabulary using the new Adaptive Dynamic Space Time Warping (A-DSTW) procedure.

The rest of the paper is organized as follows: Section 2 describes the different steps of the bottom-up sign language recognition system, including the A-DSTW algorithm. Section 3 presents the evaluation of the methodology, and finally, Section 4 concludes the paper.

2 Sign language recognition

The process for sign language recognition is shown in Figure 1. First steps of the procedure focus on segmentation of arm-hand blocks, tracking, and description of the object content, meanwhile final step uses temporal knowledge to perform sign classification.

2.1 Segmentation

In this work, we use image sequences from uncontrolled environments. In order to avoid false arm-hand detections, first, a face detection procedure based on Viola & Jones detector is applied [11]. Using the content of the detected face, a skin color model is defined [12]. This step reduces false positive detection at the same time that robustly segments arm-hand regions. Size and position of the face region are used to define a coordinate system centered on the face and normalized using the face area. The face resolution is also used to define the size of the candidate regions. This step makes the procedure invariant to scale and translation. Arm-hand regions are segmented just by capturing the highest density blobs at the expected locations. An example of this procedure is shown in Figure 2(a)-(c). First, the face region and skin candidates are shown. Next, some candidate regions over the highest density blobs are captured. Finally, an example of region tracking is shown over the input image sequence. An example of an ideal and obtained tracked sign trajectory considering both hands are shown in Figure 3.

2.2 Feature extraction

In order to describe the content of the candidate regions, we take advantage of the state-of-the-art region descriptors. In [10, 13], the authors define the feature vector $Q_{jk} = \{x_{jk}, y_{jk}, u_{jk}, v_{jk}\}$ for arm-hand candidate k at the jth frame, tracking just one arm-hand sign. x and y correspond to the spatial coordinates and u and v to the components of the movement vector. In our case, working with two arm-hand signs, the feature vector becomes $Q_{jk} = \{x^1_{jk}, y^1_{jk}, u^1_{jk}, v^1_{jk}, F^1_{jk}, x^2_{jk}, y^2_{jk}, u^2_{jk}, v^2_{jk}, , F^2_{jk}\}$, where the two super-index correspond to the left-right candidate arm-hand, and F is the HOG feature vector of the candidate region [5].

2.3 Adaptive DSTW classification

The original DTW algorithm [8] was defined to match temporal distortions between two models. Among all the defined variants of this method, in [7], the authors defined a spatio-temporal Dynamic Time Warping in order to work with a fixed number of multiple candidates. This approach has been later applied in [10, 13], where the authors defined an one hand sign recognition system. Given the high computational complexity of the DSTW approach for a high number of fixed candidates, the authors of [13] introduced the pruning of classifiers in order to reduce cost. In this section, we present an Adaptive Dynamic Space Time Warping (A-DSTW), where the number of candidates is adapted based on the arm-hand tracking process. Next, we overview the basis of the Dynamic Time Warping approaches and the A-DSTW proposal.

2.4 A-DSTW

The goal of DTW is to find an alignment warping path between two time series $Q = \{q_1, .., q_n\}$ and $C = \{c_1, .., c_m\}$. In order to align these two sequences, a $n \times m$ matrix is designed, where the position (i, j) of the matrix contains the distance between q_i and c_j. The Euclidean distance is the most frequently applied. Then, a warping path

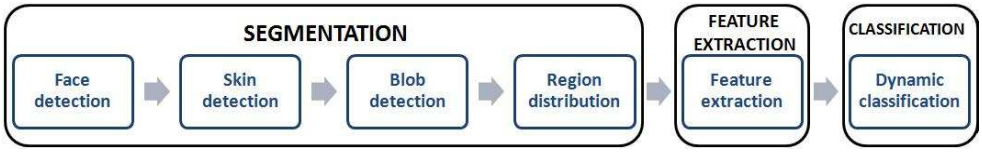

Figure 1: System scheme.

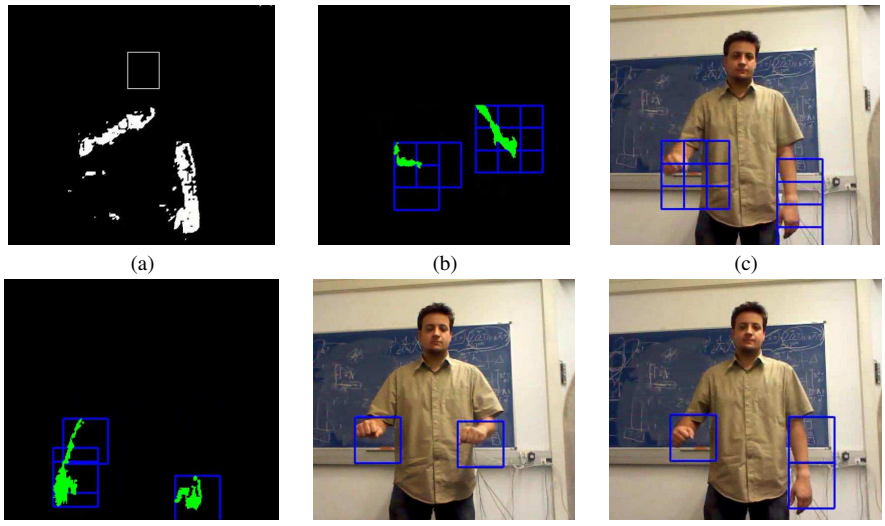

Figure 2: (a) Skin color segmentation based on face color model, (b)(c) Region distribution of DSTW for a fix number of regions over highest density blobs, and (d)(e)(f) Region distribution of A-DSTW for a variable number of regions over highest density blobs.

$W = \{w_1, .., w_T\}, \max(m, n) \leq T < m + n + 1$ is defined as a set of "contiguous" matrix elements that defines a mapping between Q and C. This warping path is typically subjected to several constraints:

Boundary conditions: $w_1 = (1, 1)$ and $w_T = (m, n)$.

Continuity: Given $w_{t-1} = (a', b')$, then $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$.

Monotonicity: Given $w_{t-1} = (a', b')$, $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$, this forces the points in W to be monotically spaced in time.

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

$$DTW(Q, C) = \min \left\{ \frac{1}{T} \sqrt{\sum_{t=1}^{T} w_t} \right\} \quad (1)$$

where T compensates the different lengths of the warping paths. This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i, j)$ as the distance $d(i, j)$ found in the current cell and the minimum of the cumulative distance of the adjacent elements: [1]

$$\gamma(i, j) = d(i, j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (2)$$

The first image in Figure 4 shows an exemple of a warping path for a two time series matched in a DTW matrix.

In the case of the DSTW of [7], the two-dimensional matrix is extended into a three-

[1] Note that though different adjacency elements can be considered varying the warping normalization factor T, here we follow the present adjacency rule as the most extended one.

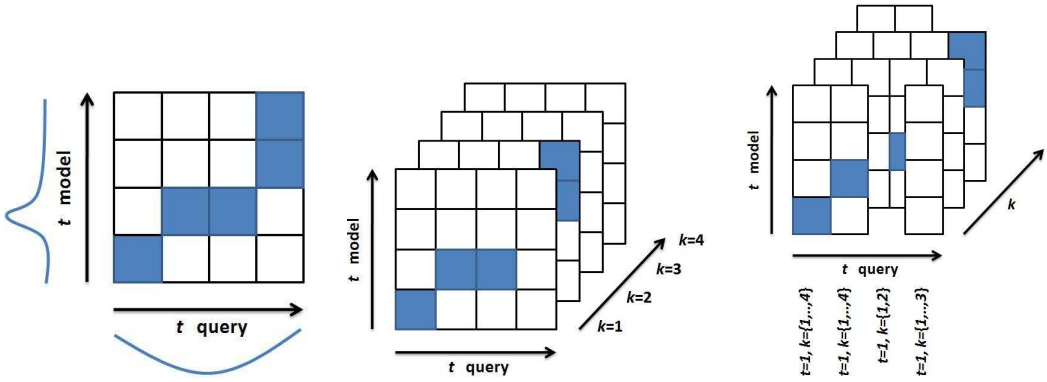

Figure 4: Examples of dynamic matching by DTW variants.

Figure 3: Left: ideal sign hand trajectories. Right: tracked hand trajectories.

dimensional one in order to match K multiple candidates of the third dimension that appear at each instant of time. An example of this procedure is shown in the second image of Figure 4. In this case, the warping path constraints are re-defined in a three-value space as follows:

Boundary conditions: $w_1 = (1, 1, k)$ and $w_T = (m, n, k')$, $k, k' \in [1, .., K]$.

Continuity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1, k, k' \in [1, .., K]$.

Monotonicity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1, k, k' \in [1, .., K]$, this forces the points in W to be monotically spaced in time.

Now, continuity and monotonicity are required only in the temporal dimensions. No such restrictions are needed for the spatial dimension; the warping path can "jump" from any spatial candidate k to any k'. In this case, the cumulative distance of the adjacent elements is re-defined as follows:

$$\gamma(i, j, k) = d(i, j, k) + \min\{\gamma((i-1, j-1), (i-1, j), (i, j-1) \times \{1, .., K\})\} \quad (3)$$

Concerning our A-DSTW, we follow the DSTW rules, but, instead of using a predefined number of candidate cases, we adapt the number of candidates based on the length of the segmented blobs. This results in a variable number of candidates per instant of time (frames in our case).

Given a sequence of model feature vectors $M_i, 1 \leq i \leq m$, and a sequence of sets of query feature vectors $Q_j = \{Q_{j1}, .., Q_{jK}\}, 1 \leq j \leq n$, where K varies among different j, now the A-DSTW warping constrains are defined as follows:

Boundary conditions: $w_1 = (1, 1, k)$ and $w_T = (m, n, k')$, $k, k' \in [1, .., \max(\text{length}(Q))]$.

Continuity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1, k, k' \in [1, .., \max(\text{length}(Q))]$.

Monotonicity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1$, $k, k' \in [1, .., \max(\text{length}(Q))]$, this forces the points in W to be monotically spaced in time.

The A-DSTW algorithm is shown in Algorithm 1. $N(w) = N(i, j, k)$ defines the set of all possible values of w_{t-1} that satisfy the warping path constraints:

$$N(i, j, k) = \{(i-1, j), (i, j-1), (i-1, j-1)\} \times \{1, .., K\} \quad (4)$$

taking into account that the value of K varies over j. An example of an A-DSTW (i, j, k) space is shown in the last image of Figure 4. Note that the number of query candidates changes across time instants. In Figure 2(d)-(f) an example of candidates distribution over segmented blob regions and original images is shown. Note that different num-

ber of candidates are distributed over the skin region based on the segmented areas.

Table 1: The A-DSTW algorithm

Input: A sequence of model feature vectors $M_i, 1 \leq i \leq m$, and a sequence of sets of query feature vectors $Q_j = \{Q_{j1}, .., Q_{jK}\}, 1 \leq j \leq n$, where K varies among different j.
Output: A global matching cost $D*$ and an optimal warping path $W* = (w_1*, .., w_T*)$.
$j = 0$ // Initialization
for $i = 0 : m$ **do**
  for $k = 1 : \max(\text{length}(Q))$ **do**
    $D(i, j, k) = \infty$
  end
end
$D(0, 0, 1) = 0$
for $j = 1 : n$ **do**
  for $i = 1 : m$ **do**
    for $k = 1 : \text{length}(Q_j)$ **do**
      if $i = 0$ **then**
        $D(i, j, k) = \infty$
      end
      else
        $w = (i, j, k)$
        $D(w) = d(w) + \min_{w' \in N(w)} D(w')$
        $b(w) = \text{argmin}_{w' \in N(w)} C(w', w)$
      end
    end
  end
end
$k* = \text{argmin}_k \{D(m, n, k)\}$ // Termination
$D* = D(m, n, k*)$
$w_T^* = (m, n, k*)$
$w_{t-1}^* = b(w_t^*)$ // Backtrack

3 Results

Before the presentation of the results, first, we discuss the data, methods and parameters, and validation protocol of the experiments.

Data: The data used in our experiments consists of 200 video sequences corresponding to 20 signs from the Spanish sign language dictionary from 10 different subjects. Half of the sequences are captured using short-sleeved clothes meanwhile the remaining half of the data is recorded using long-sleeved clothes. The resolution of the video sequences is 640×480 and 15 FPS. Some samples of the captured signs are shown in Figure 5.

Figure 5: Frame samples of the sign language database.

Methods and parameters: We use the DSTW with 15 fixed candidates per frame for comparison [10]. Concerning the A-DSTW approach, the number of candidates is adapted based on the number of regions that uniformly fall in the length of the detected blobs [14] with size 75% of the face area with an overlapping of 50% among regions. For both dynamic methods, spatial coordinates, movement vectors and HOG descriptors are computed per candidate. The weight of the four first features is of 0.5 and the same weight is assigned for the normalized HOG descriptor in the Euclidean computation in the dynamic matching.

Validation measurements: We apply stratified ten-fold cross-validation and test for the confidence interval with a two-tailed t-test. The ground truth is obtained using the samples from the ten-fold iteration containing short-sleeved clothes and tracking just one candidate region per hand. The remaining data is used for testing. Each test sample is categorized applying 3-KNN over the first three retrieved sign models from the database after computing the warping path.

3.1 Sign recognition

Applying stratified ten-fold evaluation as commented before over the sign language database for both DSTW and A-DSTW, we obtained the results shown in the top row of Table 2. A-DSTW obtains near 13% more of performance, corresponding to a relative performance improvement near

20%. This result is due to the main drawback of the DSTW algorithm. Using a fixed number of candidates, when a small arm-hand region is segmented, redundant information may be computed meanwhile when the segmented skin region is large, we may compute few candidate regions, reducing efficiency. On the other hand, simply adapting the number of candidates does not only increase the final performance, we can also save time. This can be seen by the mean number of candidate regions shown in the middle row of Table 2. In comparison to the mean of 15 fixed regions of DSTW, the A-DSTW only required a global mean of 9.75 regions, allowing a real time computation, as shown in the bottom row of Table 2.

Table 2: Sign language recognition results.

	DSTW	A-DSTW
Performance	79.27±3.15	92.18±2.12
Mean candidate regions	15	9.75
Computed frames/second	18	26

4 Conclusion

In this paper, we proposed a general framework for real time action classification applied to the sign language recognition problem. The system is based on skin blob detection and tracking. Using a coordinate system estimated from a face detection procedure, the final sign is recognized by means of a new adaptive method of Dynamic Space Time Warping. The A-DSTW uses a variable number of segmented region candidates to match temporal series, yielding a better performance while reducing the computational complexity of the classification task.

5 Acknowledgement

This work has been partially supported by the projects TIN2009-14404-C0 and CONSOLIDER-INGENIO 2010 (CSD2007-00018). We are specially grateful to Noelia H. for her support during the development of this work.

References

[1] J. Triesch, C. von der Malsburg, Robotic gesture recognition, Gesture Workshop (1997) 233–244.

[2] F. Chen, C. Fu, C. Huang, Hand gesture recognition using a real-time tracking method and hidden markov models, Image and Video Computing 21 (8) (2003) 745–758.

[3] J. Martin, V. Devin, J. Crowley, Active hand tracking, Automatic Face and Gesture Recognition (1998) 573–578.

[4] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: International Conference on Computer Vision & Pattern Recognition, Vol. 2, 2005, pp. 886–893.

[6] Y. Sato, T. Kobayashi, Extension of hidden markov models to deal with multiple candidates of observations and its application to mobile-robot-oriented gesture recognition, ICPR 2 (2002) 515–519.

[7] J. Alon, V. Athistos, Q. Yuan, S. Sclaroff, Simultaneous localization and recognition of dynamic hand gestyres, IEEE Motion Workshop (2005) 254–260.

[8] J. Kruskall, M. Liberman, The symmetric time warping algorithm: From continuous to discrete, Time Warps. Addison-Wesley.

[9] H.-D. Yang, S. Sclaroff, S.-W. Lee, Sign language spotting with a threshold model based on conditional random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (7) (2009) 1264–1277.

[10] A. Stefan, V. Athitsos, J. Alon, S. Sclaroff, Translation and scale-invariant gesture recognition in complex scenes, in: PETRA: Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments, 2008, pp. 1–8.

[11] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2001, pp. 511–518.

[12] M. Jones, J. Rehg, Statistical color models with application to skin detection, in: Intenational Journal of Computer Vision, Vol. 46, 2002, pp. 81–96.

[13] J. Alon, V. Athistos, Q. Yuan, S. Sclaroff, A unified framework for gesture recognition and spatiotemporal gesture segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (9) (2009) 1685–1699.

[14] O. B. segmentation software, http://opencv.willowgarage.com/wiki/cvblobslib.