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
Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain

Centre de Visió per Computador

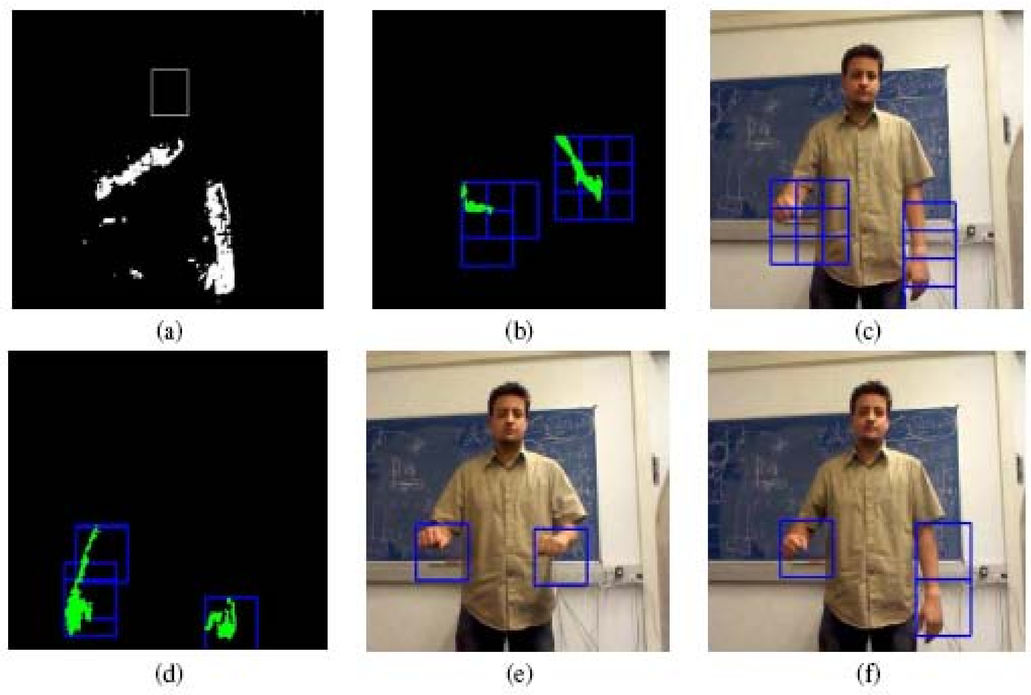
ABSTRACT

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.

1.SEGMENTATION & FEATURE EXTRACTION

In this work, we use image sequences from uncontrolled environments. In order to avoid false armhand detections, first, a face detection procedure based on **Viola & Jones detector** is applied. Using the content of the detected face, a **skin color model** is defined. 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 region**s. 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.

In order to describe the content of the candidate regions, we take advantage of the state-of-the-art region descriptors. In other works, the authors define the **feature vector** $Q_{jk} = \{x_{jk}, y_{jk}, u_{jk}, v_{jk}\}$ for armhand 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 $Q_{jk} = \{x_{jk}^1, y_{jk}^1, u_{jk}^1, v_{jk}^1, F_{jk}, {}^2x_{jk}, {}^2y_{jk}, {}^2u_{jk}^2, v_{jk}^2, F_{jk}\}$, where the two super-index correspond to the left-right candidate arm-hand, and F is the **HOG feature vector** of the candidate region



(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.

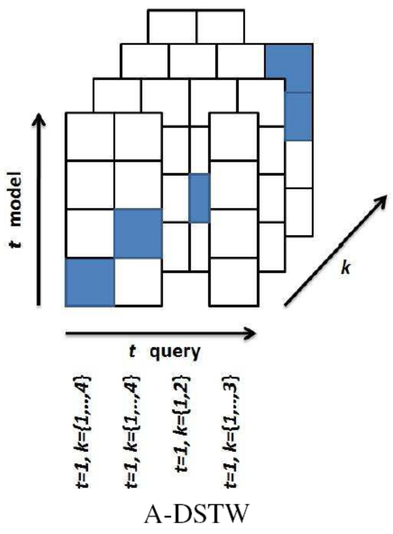
2. A-DSTW

In our work, we define a sequence of model feature vectors M_i; $1 \le i \le m$, and a sequence of sets of query feature vectors $Q_j = \{Q_{j1},..,Q_{jK}\}$, $1 \le j \le n$, where K varies among different j, 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(length(Q))].
- *Continuity:* Given $w_{t-1} = (a'; b'; k')$, then $w_t = (a, b, k)$, $a - a' \le 1$ and $b - b' \le 1$, $k,k' \in [1,..,$ max(length(Q))].
- *Monotonicity:* Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \ge 0$ and $b - b' \ge 0$, $k,k' \in [1,..,$ max(length(Q))], 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:

$$A\text{-}DSTW(M,Q) = \min\left\{\frac{1}{T}\sqrt{\sum_{t=1}^{T} w_t}\right\}$$

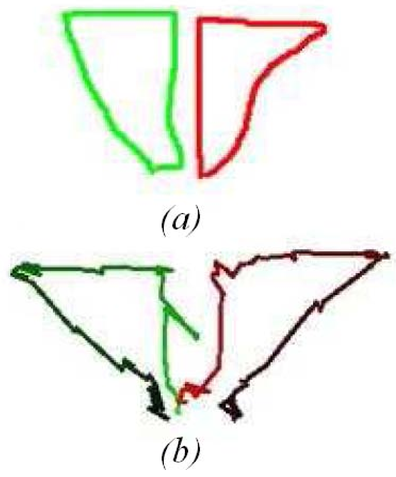


A-DSTW

3.RESULTS

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 the table. A-DSTW obtains near **13% more of performance**, corresponding to a relative performance improvement near 20%. 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 the table.

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



(a)Ideal hand trajectories
(b)Tracked hand trajectoires

4. CONCLUSIONS

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.