Automatic Sign Language Recognition: vision based feature extraction and probabilistic recognition scheme from multiple cues

George Caridakis
Image, Video and Multimedia Systems Lab
National Technical University of Athens
Iroon Polytexneiou 9, 15780
Athens, Greece
gcari@image.ntua.gr

Olga Diamanti
Computer Vision, Speech Communication and Signal Processing Group
National Technical University of Athens
Iroon Polytexneiou 9, 15780
Athens, Greece
olga.diam@gmail.com

Kostas Karpouzis
Image, Video and Multimedia Systems Lab
National Technical University of Athens
Iroon Polytexneiou 9, 15780
Athens, Greece
kkarpou@cs.ntua.gr

Petros Maragos
Computer Vision, Speech Communication and Signal Processing Group
National Technical University of Athens
Iroon Polytexneiou 9, 15780
Athens, Greece
maragos@cs.ntua.gr

ABSTRACT

This work focuses on two of the research problems comprising automatic sign language recognition, namely robust computer vision techniques for consistent hand detection and tracking, while preserving the hand shape contour which is useful for extraction of features related to the handshape and a novel classification scheme incorporating Self-organizing maps, Markov chains and Hidden Markov Models. Geodesic Active Contours enhanced with skin color and motion information are employed for the hand detection and the extraction of the hand silhouette, while features extracted describe hand trajectory, region and shape. Extracted features are used as input to separate classifiers, forming a robust and adaptive architecture whose main contribution is the optimal utilization of the neighboring characteristic of the SOM during the decoding stage of the Markov chain, representing the sign class.

Categories and Subject Descriptors

I.4.6 [**Computing Methodologies**]: Segmentation; I.5.4 [**Computing Methodologies**]: Pattern RecognitionApplications

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Copyright 2008 ACM XXXXXXXXXĚ $5.00.
PETRA 2008, July 15-19,2008, Athens, Greece.

General Terms

Algorithms, Languages, Human Factors, Design

Keywords

Sign Language Recognition, Gesture Recognition, Self-Organizing Maps, Markov Models, Hidden Markov Models, Image Segmentation, Geodesic Active Contour Models, Optical Flow, Feature Extraction

1. INTRODUCTION

Sign languages is correctly considered as the top of the gesture hierarchical taxonomy. The importance of such a group of languages establishes the automatic recognition of sign languages as a research challenge for various disciplines including computer vision, machine learning, human action understanding and natural language processing. Sign language is the least standardized, largely symbolic and referential, highly structured amongst the gesture classes. Features comprising co-articulation of several signals such as hand/arm gestures, facial expressions, head movements, body postures and torso movements makes the task of recognizing isolated or continuous signing a highly complex one and although a large number of approaches have been proposed, robust automatic sign language recognition still remains an open problem.

There are several factors that impede the task of automatic sign language recognition. Sign languages are highly inflected, resulting in too many appearances of inflectional variants to model them all separately. Many signs can be modified according to some grammatical function, such as number, subject-verb agreement, and verb-object agreement. They can also be modified to indicate aspect (e.g., fast,

slow), repetition, and duration. Furthermore, events occur both sequentially and simultaneously. Unlike speech recognition, sign language recognition cannot consider all possible combinations of simultaneous events explicitly, because of their sheer number.

The structure of this paper is as follows. A brief overview of work in the broad research area of automatic sign language recognition is presented in section 2. Section 3 describes the adopted methodology, both at the level of computer vision techniques for hand detection, tracking and feature extraction (section 3.1) and at the level of recognition employing different models for various streams of information (section 3.2) and fusing the outputs at the decoding stage. Finally, section 4 summarizes the proposed architecture and future work is presented.

2. RELATED WORK

An extensive review of several techniques is presented both in [7] and [15]. The first focuses mainly on SL recognition and classification issues, while examining closely hand localization and tracking, and on various feature extraction techniques related to automatic analysis of manual signing. In addition, it addresses the linguistic aspect of SL and non manual signals, along with methodologies to incorporate these in the SL recognition chain. On the other hand, Wu and Huang delve more into works related to hand modeling (shape analysis, kinematics chain and dynamics) and computer vision, and pattern recognition issues associated to hand localization and feature extraction from image sequences. Classification schemes involve several methods, depending on the features and the stages of the procedure. Methods used include neural networks and variants, hidden markov models and variants, principal component analysis, and numerous other machine learning methods or combinations (decision trees, template matching, etc.).

One of the most commonly proposed approaches involves feature extraction from the input signal and utilization of these features as input for a fine tuned HMM [11]. In addition, variations of the previous group have been widely adopted [9], [13]. Other approaches employ alternate machine learning and artificial intelligence techniques such as recurrent fuzzy network , time delay neural network [16], finite state machines [12], Bayesian classifiers [14], etc. Finally, there have been several efforts combining more than one technique. Mantyla et al. [6] present a system for static gestures recognition using a self-organizing mapping scheme, while a hidden Markov model is used to recognize dynamic gestures. Black and Jepson [1] present an extension of the "condensation" algorithm, modeling gestures as temporal trajectories of the velocity of the tracked hands. Fang et al. [4] present an additional layer enhancing the HMM architecture with SOFM and improving their recognition rate by 5%, while introducing a fuzzy decision tree in an attempt to reduce the search space of recognized classes without loss of accuracy.

3. PROPOSED ARCHITECTURE

The overall system consists of two main modules, image processing and classification. Geodesic Active Regions models enhanced with color and motion cues evolve, minimizing an error function, to fit the hand regions and features relevant to hand location, region and shape are extracted. The latter are used as inputs for classifiers based on hand location, motion direction, region based features, Fourier descriptors, shape moments and curvature coefficients. The outputs of the classifiers are then fused accordingly and a final classification decision is made.

3.1 Computer Vision

3.1.1 Hand detection and tracking

In order for a feature vector to be extracted from the sign language videos, a stage of visual processing is required. In this stage, each video frame is segmented in order to isolate the signer's hands, from which the relevant information can be extracted. For the segmentation of the video frames we shall use the geodesic active regions model, introduced by Paragios et al in [10], and based on the geodesic active contour (GAC) model proposed by Caselles et al. in [2]. The GACs are deformable two-dimensional contours, which evolve to minimize a suitable energy function, designed to meet the specific needs of the segmentation process. The process results in a robust and reliable hand detection and tracking as can be shown in figure 3.1.1.

The Geodesic Active Region (GAR) model. Let C be a planar curve with arclength parameter s and length $L(C)$, and let $\vec{C}(s) = [x(s), y(s)] : [0, L(C)] \rightarrow \mathbb{R}^2$ be its arc-length parametrization. In the GAC model we aim to minimize the function $E = \int_0^{L(C)} g(I(\vec{C}(s)))ds$, where I is the intensity image we wish to segment. The function $g : [0, +\infty] \rightarrow \mathbb{R}^2$ is a stopping function, designed to assume minima at image edges, with the property $g(r) \rightarrow 0$ as $r \rightarrow +\infty$. It can be proven that the selected energy function ensures that the stable state of the curve will satisfy some smoothness criteria and will also tend to locate itself in regions of the image where the image gradient magnitude is relatively large (i.e. on the image edges). The minimization of the energy function is conducted by means of the steepest descent method, resulting in an Euler-Lagrange PDE for the evolution of the curve.

In order for the numerical solution of this PDE to allow topological changes to the curve, the GAC model is usually combined with the level-sets method, introduced by Osher and Sethian in [8]. In the level-set framework, the contour C is defined implicitly as the zero level set, at each time step, of an embedding scalar function u defined on the image plane: $C(t) = \{(x, y) : u(x, y, t) = \lambda\}$. A commonly used embedding surface is the signed distance function from the evolving contour. Once we have defined the contour in terms of the embedding function, we can extend the evolution PDE for the contour to obtain the evolution PDE for the function u:

$$\frac{\partial u}{\partial t} = div\left(g(I)\frac{\nabla u}{\|\nabla u\|} + F(u)\right)\|\nabla u\| \qquad (1)$$

The GAC model can be enhanced with the addition of external forces, represented by $F(u)$ in eq.(1), to the evolution PDE. An example is the Geodesic Active Region model ([10]), in which the image is partitioned into two or more regions, which are assumed to be homogenous with respect to

some particular statistically modeled image feature. When the image consists of only two regions, A and A^c, we obtain the following equation for the evolution of u:

$$\frac{\partial u}{\partial t} = div\left[g(I)\frac{\nabla u}{\|\nabla u\|} + \alpha \log\left(\frac{P_A(I)}{P_{A^c}(I)}\right)\right]\|\nabla u\| \quad (2)$$

where $P(A)$ denoted the probability of pixel x belonging to region A, based on the statistical model for this region. In the above equation, which we will extensively use throughout the paper, the evolution is guided by a region based force and an edge based force.

Combining skin color information with the GAR model. Skin color segmentation is feasible because the human skin has a color distribution that usually differs from that of the background. The intensity image I can thus be partitioned into two separable regions, one being the union of the skin-colored regions, and the other consisting of the rest of the image pixels, which will be referred to as "background". We may therefore adapt the geodesic active region model to introduce a new force for skin segmentation:

$$F_{color} = \log\left(\frac{P_s(\vec{x})}{P_b(\vec{x})}\right) + cg(I) \quad (3)$$

where P_s, P_b denote the probability of a certain pixel belonging to the skin or background regions, respectively. The above force consists of the region-based statistical color force and a second force, known as "balloon force" ([2]) which speeds up the evolution procedure by attempting to minimize the surface of the embedded curve.

We estimate the probability P_s via a skin color model. This model is constructed by utilizing the color information, in terms of the (a, b) color coefficients in the Lab color space, of several skin colored regions cropped from real input images. The background probability is derived straightforwardly as $P_b = 1 - P_s$. Our proposed force ensures that the curve will eventually converge to those image edges that separate skin regions from the background. The use of the geodesic active region framework eliminates any issues concerning the continuity of the skin regions detected by the color model, as the smooth curve will enclose the whole skin region, provided that these discontinuities are not too large. On the other hand, the statistical force makes the segmentation model more robust with respect to weak or false intensity edges.

Incorporation of motion information. While color information is a major cue that assists in the detection of hands in images, motion information is equally crucial when the goal is to recognize human gestures. The employment of motion information in the segmentation module may prove to be extremely effective in such cases; as far as sign language recognition is concerned, it may help to resolve the well-known problems arising in the presence of hand-face or hand-hand occlusions, given that the signer's face remains relatively motionless while the hands move vividly. In this work we exploit the available motion information by using the geodesic active region model, in a way similar to the skin color segmentation described in the section 3.1.1. Namely, we will again use a statistical force of the logarithmic form presented in previous section, with the image at hand now being partitioned into two regions, the one comprising of the moving pixels and the other of the pixels whose position remains the same.

The motion information is provided from the optical flow field, derived with the well-known Lucas-Kanade algorithm ([5]). The magnitude of the optical flow field could be used to obtain the probability of a certain image pixel belonging to either the static or the moving component. Thus, we introduce a new evolution force for the active contour, according to the following steps:

1) Estimation of the optical flow field OF, its magnitude $|OF(x,y)| = \sqrt{V_x^2 + V_y^2}$ and the moving region probability $P_{mov}(x,y) = |OF(x,y)|/\max(|OF|)$,

2) Motion force: $F_{mov}(x,y) = \log\left(\frac{P_{mov}(x,y)}{P_{stat}(x,y)}\right)$

where $P_{stat}(x,y) = 1 - P_{mov}(x,y)$ is the probability of the pixel at location (x,y) belonging to the static region ("background").

This new motion force operates in an analogous way to the color force. It leads the evolving contour to converge so as to include regions, where motion is detected. This could be used to locate the hands in an image and discriminate them from the face region, which will also be detected by the skin color model. Thus, we can diminish or eliminate errors in the estimation of the hands' positions in the presence of occlusions. The overall force at point $\vec{x}$ is:

$$F(\vec{x}) = \log\left(\frac{P_{skin}(\vec{x})}{P_{nonskin}(\vec{x})}\right) + \log\left(\frac{P_{mov}(\vec{x})}{P_{stat}(\vec{x})}\right) + cg(I(\vec{x})) \quad (4)$$

Figure 1: Image processing segmentation and tracking results

3.1.2 Feature extraction

A variety of features can be used for sign language recognition. In this work we use features that primarily describe the shape of the segmented signer's hands, in order to represent the handshapes used by the signer, which are the main source of information with regard to the interpretation of a specific sign. Motion information is also extracted by following the trajectory of the hands' centroids, as will be described in the following sections.

As shape features we use a variety of descriptors, both boundary-based (Fourier descriptors, Curvature features) and region-based (Moments, Moment-Based Features).

Fourier Descriptors. The extraction of the Fourier descriptors for a given shape is based on the notion that any digital shape boundary can be represented by a periodic complex function $z_i = z(i) = x_i + jy_i$, where x_i, y_i, $i = 0, 1, ..., N-1$ are the horizontal and vertical position coordinates of each of the N boundary points. Such a function can be transformed into a Fourier series

$$Z_k = \sum_{n=0}^{N-1} z_n e^{-2\pi jkn/N} \ , \ k = 0, 1, 2, \ldots, N-1 \qquad (5)$$

The Fourier descriptors are derived straightforwardly from the coefficients Z_k of the above Fourier series. Namely, the coefficients Z_0, Z_1 are ignored and the rest of the coefficients are divided by $|Z_1|$, yielding the Fourier descriptors C of the shape

$$C_{k-2} = \frac{\|Z_k\|}{\|Z_1\|} \ , \ k = 2, 3, \ldots, N-1 \qquad (6)$$

It can be easily proven that this form of the coefficients ensures that the resulting shape decriptors remain unaffected by shape translations, rotations, scalings and changes of the starting point on the boundary.

Curvature Cepstrum Coefficients. This set of shape decriptors are based on the computation of the cepstrum of the shape's curvature, in a way directly analogous to the extraction of the MFCC features from voice signals. Namely, the first stage involves the extraction of the curvature from the binary shape of the hand, by means of Freeman's chain code method. Next, the cepstrum of the curvature signal is extracted, and the largest N_C coefficients are selected and included in the feature vector. Extensive experimentation has shown that the curvature function can be sufficiently reconstructed by a fairly small set of cepstrum coefficients, thus obtaining a compression of the original signal length by over 90%.

Shape Moments. Invariant shape moments are also extensively used in shape representation and recognition, since they achieve significant data compression. The pq^{th}-order central (i.e., with respect to the center of mass (μ_x, μ_y)) moment of a binary shape I is defined by the following equation:

$$\mu_{ij} = \sum_{x=1}^{N} \sum_{y=1}^{N} (x - \mu_x)^i (y - \mu_y)^j I(x, y) \qquad (7)$$

thus yielding the normalized central moments

$$\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{\gamma}}, \quad \gamma = \frac{i+j}{2} + 1 \qquad (8)$$

The above moments can be used to obtain a set of seven scaling, translation and rotation invariant measures, given by the following equations:

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{02} - \eta_{20})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} + \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2\right] + \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{03} + \eta_{21}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2\right] \\
\phi_6 &= (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2\right] + \\
&\quad + (\eta_{20} - \eta_{02}) \left[4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21})\right] \\
\phi_7 &= (3\eta_{21} + \eta_{03})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12}) - 3(\eta_{03} + \eta_{21})^2\right] + \\
&\quad + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]
\end{aligned} \qquad (9)$$

Other Region-Based Features. In the feature vector we also added a set of other region-based features, related to the moments described above. These include: the **area** of the shape, its **eccentricity**, its **compactness**, its **minor and major axis lengths**, and its **orientation**.

3.2 Sign language Recognition

Sign recognition is performed by fusing separate component models for sign trajectory and hand shape cues. A novel approach is introduced by applying a combination of self organizing maps and markov models for sign trajectory classification. The extracted features used in the trajectory module include the trajectory of the hand and the direction of motion in the various stages of the gesture. This classification scheme is based on the transformation of a sign representation from a series of coordinates and movements to a symbolic form and on building probabilistic models using these transformed representations. Concerning hand shape, Hidden Markov Models are used to classify each sign instance into one of the models created by training a unique model for every corresponding class. Our study indicates that, although each of the two sets of features (trajectory and hand shape information) can provide distinctive information in most cases, only an appropriate combination can result in robust and confident user independent sign language recognition.

The steps of the introduced procedure, which is depicted in figure 2, begin with the image processing module described in section 3.1. Following, each isolated sign instance is represented by a time series of points, representing the hand's location with respect to the head of the signer and a set of features aiming to describe the distinct handshapes. Consequently, a sign G_i containing l points can de expressed as an ordered set of points:

$$G_i = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \qquad (10)$$

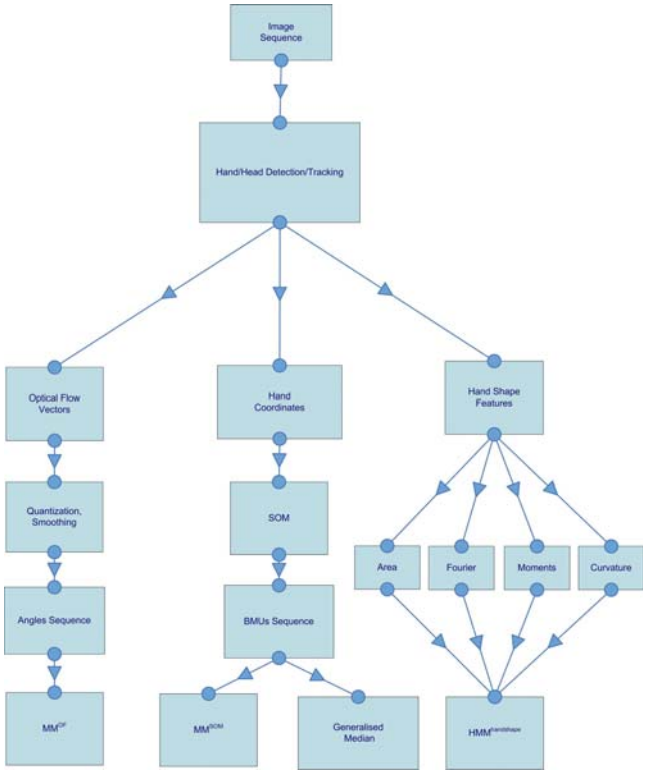

Figure 2: System architecture

$$
\begin{aligned}
HS_i &= \{HS_i^a HS_i^f HS_i^m HS_i^c\} \\
HS_i^a = HS_i^{area} &= \begin{bmatrix} HS_{i_{1..l}}^{region} \\ HS_{i_{1..l}}^{eccentricity} \\ HS_{i_{1..l}}^{orientation} \\ HS_{i_{1..l}}^{ratio} \\ HS_{i_{1..l}}^{compactness} \end{bmatrix} \\
HS_i^f &= HS_i^{fourier} = \{HS_{i_{1..l}}^{C_{1..20}}\} \\
HS_i^m &= HS_i^{moments} = \{HS_{i_{1..l}}^{\phi_{1..7}}\} \\
HS_i^c &= HS_i^{curvature} = \{HS_{i_{1..l}}^{N_{1..31}}\}
\end{aligned} \tag{11}
$$

where l varies across different sign classes. The system's input is a set of sign instances D, assigned to c different sign categories.

The proposed modeling scheme is based on the transformation of a sign representation from a series of coordinates and movements to a symbolic form which, in turn, is used to build the respective probabilistic models. The first transformation is based on the relative position of the hand during the sign and is achieved using a self-organizing map model. Despite the fact that the map units are treated as symbols, the map's neighborhood function provides a distance metric between them, that is used during the classification of an unlabeled gesture. Additionally, this enables the use of the Levenshtein distance metric for the comparison between these sequences of symbols and the definition of a 'mean' string of symbols representing e.g. the signs included in a D_j set.

An additional transformation is based on the optical flow of the gesture, aiming to describe the hand direction changes during signing. The symbols generated from this transformation constitute the set of angles of the hand's trajectory. This set is limited to quantized values that are treated as symbols in order to be used for the creation of an additional set of Markov models.Furthermore, hand shape features aiming to describe both the hand configuration and the palm orientation are used to train continuous HMMs Gaussian mixture components.

For the classification of an unlabeled sign instance, all the above mentioned trained models are tested against this instance and participation probabilities are fused, balanced using weights calculated according to the isolated recognition rates, thus achieving a robust recognition scheme tackling cases of low confidence or ambiguity.

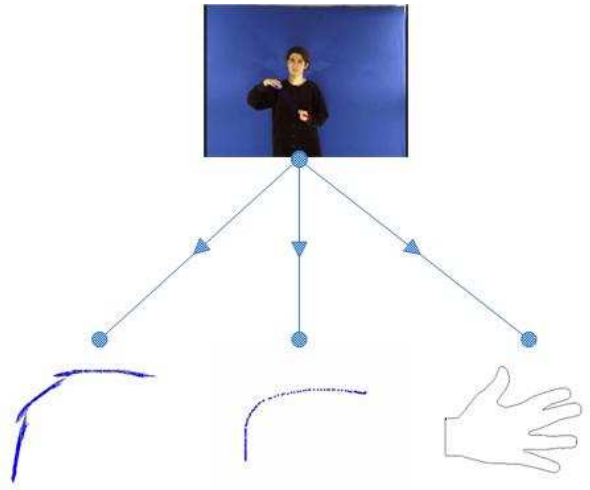

Figure 3: A more intuitive system architecture overview

3.2.1 Building Sign Models

The coordinates of all the points from all the gestures are used to train a hexagonal, two-dimensional grid SOM with the batch mode learning procedure. The points are fed to the map in an unordered form, inconsequently to the gesture instance they belong to and to their ranking position into the gesture. Following training, each point is assigned to the respective best matching unit (BMU) on the map, i.e. the unit of the map closer to the point in the input data space, according to the Euclidean distance of the two vectors. Thus, a gesture G_i can be transformed from a series of points to a series of map units.

$$
T(G_i) = (u_1, u_2, \ldots, u_l) \colon u_i = BMU(x_i, y_i) \tag{12}
$$

Function $BMU(x_i, y_i)$ returns the index of the best-matching unit for point (x_i, y_i) and $T(G_i)$ is the modified gesture representation. Given that u_i is the index of a map unit, this function can be is declared as $BMU : R^2 \rightarrow S$, where S is the set of the indices of all map units and can be treated as a set of symbols. In many cases, the u_i value of consequent points of a gesture remains the same since, although the continuous movement of the hand is represented by the distinct points, consequent points are generally close in the input data space. Replacing consequent equal values of u_i with a single value results in the following gesture definition,

$$
\begin{aligned}
G_i^{'} = N(T(G_i)) &= \{u_1, u_2, \ldots, u_m, \} \\
&: m \leq l, u_t \neq u_{t-1} \forall t \in [2, l]
\end{aligned} \tag{13}
$$

where N is a function that removes consecutive equal u_i values and $G_i^{'}$ is the transformed gesture instance. The transformation of the gestures with the use of the SOM can be considered a transformation of the continuous trail to a sequence of m discrete symbols, different for every gesture class, that define the finite states to build first order Markov chain models.

Such a model, for each of the categories in the gestures' data set, is created. The sequence of the u_i values into the transformed gestures $G_i^{'}$ of $D_j^{'}$ set, will be used for the calculation of the transition probabilities of the model MM_j^{som} describing the j category and for the determination of the values of the function π_j^{som}, which is the first state probability function of this model. The result is a set MM^{som} of c Markov models.

$$\begin{aligned} MM^{som} &= \{MM_1^{som}, MM_2^{som}, \ldots, MM_c^{som}\} \\ &: D_i^{'} = \{G_1^{'}, G_2^{'}, \ldots, G_n^{'}\} \rightarrow MM_i^{som} \end{aligned} \tag{14}$$

These models are used to evaluate a new unlabeled gesture in order to be classified in one of the c categories. Figure 4 depicts the above described transformation for a gesture instance.

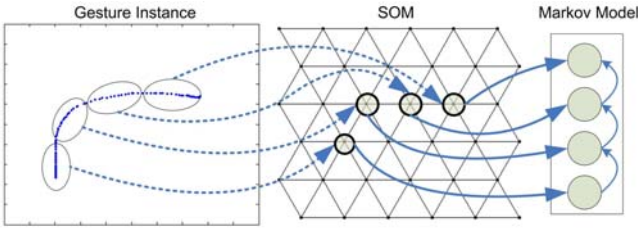

Figure 4: Correspondence of gesture trajectory points to their respective BMUs on the SOM. These BMUs constitute the states of the Markov models.

With the purpose of providing a more descriptive representation of each gesture instance, an additional transformation is introduced, based on the optical flow of each gesture. This describes the different directions that the gesture trajectory presents instead of the spatial position of gesture points. In order to achieve such a representation, direction vectors are calculated from the consecutive gesture trajectory points. These angles are then quantized in 8 different symbolic values as depicted in figure 3.2.1. The segments of coordinates in figure 4 and 3.2.1 are considered to be a set of coordinates that belong to the same cluster (BMU and Quantized Angle for figure 4 and 3.2.1 respectively). In that sense, we define the transformation of a gesture instance G_i using the OF function as:

$$\begin{aligned} OF(G_i) &= \{v_1, v_2, ..., v_m\} \\ &: v_i = W_r(Q(\arctan(\frac{y_i - y_{i-1}}{x_i - x_{i-1}}))) \end{aligned} \tag{15}$$

where v_i are the quantized values, Q the quantization function and W_r a median function applied to the values of a fixed length window around the input value. The purpose of the later is to smooth the quantized values against possible instabilities of the hand during the gesture. Applying the transformation function along with function N (equation 13) for the removal of the equal consecutive values we get

$$G_i^{''} = N(OF(G_i)) = \{v_1, v_2, ..., v_m\} \tag{16}$$

The v_i values define the states for a new set of Markov models MM^{of} that is built using the transformed set $D_j^{''}$. The first state probability function π_j^{of} is also calculated using this set.

$$\begin{aligned} MM^{of} &= \{MM_1^{of}, MM_2^{of}, ..., MM_c^{of}\} \\ &: D_i^{''} = \{G_1^{''}, G_2^{''}, ..., G_n^{''}\} \rightarrow MM_i^{of} \end{aligned} \tag{17}$$

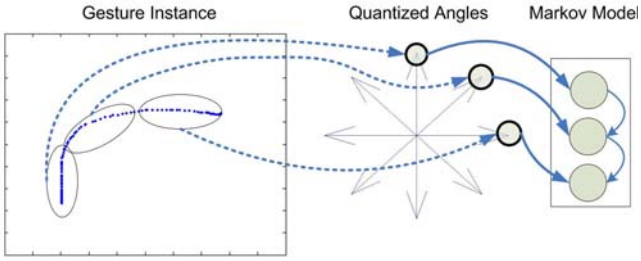

Figure 5: Building a Markov model for a gesture's optical flow

Additionally we train 4 continuous (mixture of three Gaussian), left-to-right Hidden Markov Models with the training set being features describing the handshape. As can be shown in figure 3.2.1 features describing the area of the extracted hand HMM^{hs_1}, Fourier descriptors HMM^{hs_2}, moments HMM^{hs_3} and coefficients of the Curvature Cepstrum HMM^{hs_4} are utilized to model different combinations of finger joint angles and palm orientation.

3.2.2 *Sign Decoding*

The classification of an input gesture will be based on the two sets of Markov models (equations 14 and 17). Let G_k be a gesture instance of unknown category, and $G_k^{'}$ and $G_k^{''}$ its transformed representations. Using the MM^{som} set of models, the probability of this gesture to belong in category j can be calculated as:

$$P(G_k^{'}|MM_j^{som}) = \frac{\sum_{i=1}^{m} S_i^{som}}{m} \tag{18}$$

The above equation averages the values S_i^{som}, which represent an evaluation factor for each u_i value of the $G_k^{'}$ transformed gesture with respect to the MM_j^{som} Markov model. These values are calculated as:

$$S_i^{som} = \max_z(NF_{u_i}^{som}(z)P(z|u_i, MM_j^{som})) \tag{19}$$

$$u_i = \arg\max_z(S_i^{som}) \tag{20}$$

where z is a variable that indexes the units of the trained map, $NF_{u_i}^{som}(z)$ is the distance of the unit z as defined by the self-organizing map Gaussian neighborhood function with the u_i unit as its center. In equation (9), the proximity between the state-unit z and the previous state-unit u_{t-1} of

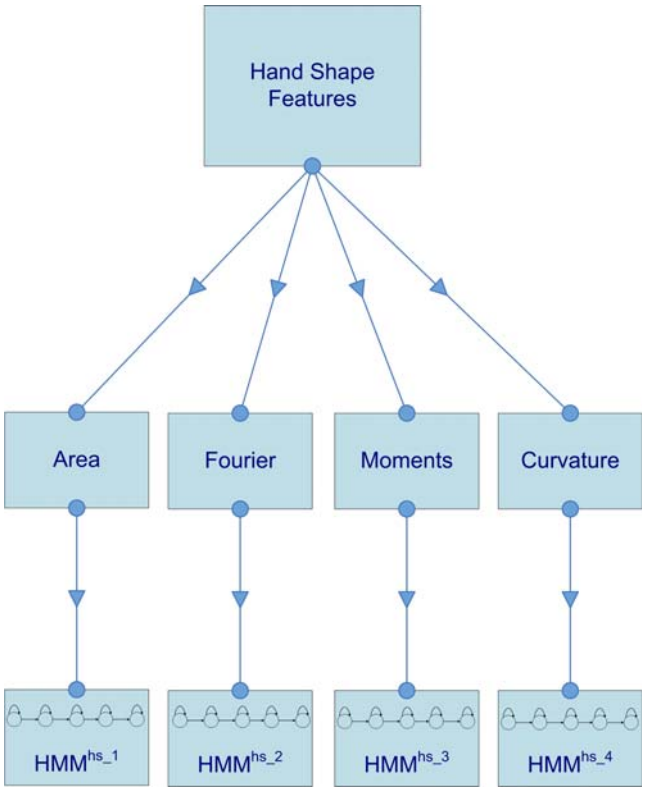

Figure 6: Hidden Markov Models based on features describing the handshape

the gesture is multiplied with the probability of the transition from state-unit z to state-unit u_{t-1}. As the z variable varies across all the units of the map, this product will provide the unit that combines a considerable transition probability from the previous state with a small distance onto the map grid from the current state. This unit will also be used as the previous state in the next step as defined by equation 20. The initial values used in the sum derive from the following equations.

$$\begin{aligned} S_1^{som} &= \max_z(NF_{u_1}^{som}(z)\pi_j^{som}(z)) \\ &: u_1 = \arg\max_z(S_1^{som}) \end{aligned} \tag{21}$$

Using the MM^{of} set of models, the probability of this gesture to belong in category j can be calculated as:

$$P(G_k^{''}|MM_j^{of}) = \frac{\sum_{i=1}^{m} S_i^{of}}{m} \tag{22}$$

The values S_i^{of} are calculated from the following equations:

$$\begin{aligned} S_i^{of} &= \max_z(NF_{v_{i-1}}^{of}(z)P(z|v_{i-1}, MM_j^{of})) \\ &: v_i = \arg\max_z(S_i^{of}) \end{aligned} \tag{23}$$

where z is a variable that indexes the different states-directions and $NF_{u_i}^{of}(z)$ a distance function between these states. These equations implement a search similar to the previous search on the map grid, but in this case the search is performed among the different possible gesture directions. The initial values are calculated in a similar way from the following equations.

$$\begin{aligned} S_1^{of} &= \max_z(NF_{v_1}^{of}(z)\pi_j^{of}(z)) \\ &: v_1 = \arg\max_z(S_1^{of}) \end{aligned} \tag{24}$$

In order to compare the length of the unknown gesture with the length of the gestures included in each $D_j^{'}$ set, a distance metric for the comparison of symbol strings is necessary. From each set $D_j^{'}$, a *Generalized Median* gesture is calculated. Let S be a set of symbol strings s_i. We can then define m as a string that consists of a combination of all or some of the symbols used in the set and which minimizes the following expression.

$$\sum_{s_i} L(s_i, m), \forall s_i \in S \tag{25}$$

where L, denotes the Levenshtein distance, one of the most widely used string distance metric. If the search for string m is restricted to the members of the set then m is the *set median*. But if m is a hypothetical string and the search is not restricted then m is the *Generalized Median* of the set. Using the above definition we calculate the Levenshtein distance $L_{kj} = L(G_k^{'}|M(D_j^{'}))$ between $G_k^{'}$ and the *Generalized median* $M(D_j^{'})$ of each $D_j^{'}$ set.

The category of the unknown gesture is primarily decided using the MM^{som} set of models. Subsequently, the category would be equal to:

$$\arg\max_j P(G_k^{'}|MM_j^{som}) \tag{26}$$

In order for the category of the unknown gesture to be decided by the above equation the four following conditions must be fulfilled.

$$\max_j(P(G_k^{'}|MM_j^{som})) \geq \alpha \tag{27}$$

$$\max_j(P(G_k^{'}|MM_j^{som}) - 2^{nd}\max_j(P(G_k^{'}|MM_j^{som}) \geq \beta \tag{28}$$

$$L_{k,\arg\max_j(P(G_k^{'}|MM_j^{som})} \leq \gamma LM(\arg\max_j(P(G_k^{'}|MM_j^{som})) \tag{29}$$

$$\max_j(\prod_{i=1}^{4}\prod_{q\in[a,f,m,c]} P(HS_k^q|HMM_j^{hs_i})) \geq \delta \tag{30}$$

The two first conditions require that the maximum probability calculated using position based models must exceed

a threshold value a, while the difference between the maximum probability and the second ranked ones must also exceed a threshold value β. These two values represent confidence thresholds. The third condition applied is that the Levenshtein distance between the gesture and the *Generalized Median* of the category with the maximum probability must be larger than the LM value of this category, multiplied by a user defined factor γ. This comparison is made in order to assess the length of the unknown gesture with respect to the average length of the gestures of the category with the maximum probability. The last criterion denotes that the product of the log-likelihood probabilities of the sign instance against four types of HMM models, namely area model HMM^{hs_1}, Fourier descriptors model HMM^{hs_2}, Moments model HMM^{hs_3} and Curvature Cepstrum coefficients model HMM^{hs_4}, should uphold a minimum value δ. If one of these conditions is not fulfilled then the category of the unknown gesture is defined from a combination of values:

$$\begin{aligned} &\arg\max_j (P(G_k^{'}|MM_j^{som})P(G_k^{''}|MM_j^{of})\frac{1}{\frac{L_{kj}}{\|M(D_j)\|}}\Pi) \\ &: \Pi = \prod_{i=1}^{4}\prod_{q\in[a,f,m,c]} P(HS_k^q|HMM_j^{hs_i}) \end{aligned} \tag{31}$$

This classification rule combines the evaluation provided from both the MM^{som} and MM^{of} set of Markov models with the Levenshtein distance of the gesture and the *Generalized median* of the each category normalized by the length of the *Generalized median* and the HMM^{hs_m} models trained with different feature sets describing the handshape.

4. CONCLUSIONS

We propose an original automatic sign language recognition architecture which consists of robust computer vision techniques for consistent hand detection and tracking, feature extraction of related to the hand location, shape, and region and a novel classification scheme incorporating Self-organizing maps, Markov chains and Hidden Markov Models. Extracted features train separate classifiers, which in turn are fused into a decision level, committee-machine-like setup, during the classification stage, enhancing the proposed architecture with multimodality and robustness against noisy and unconstrained environments or sign inflection.

Ongoing work includes the overall validation of the proposed scheme, since the effectiveness of each component, forming the recognition process, has already been tested. This validation will use as a corpus the one constructed by Efthimiou and Fotinea in [3]. Furthermore we are willing to compare the recognition rates against other approaches found in the literature, either by implementing them or by applying the proposed architecture on other corpora. Finally, we will investigate ways to extend the platform to perform sign language recognition on a sign-component level (signeme) and on continuous signing.

5. ACKNOWLEDGMENTS

This work has been partially financed by the European Union in the framework the the National Programme - Information Society, Measure 3.3: 'Image, Sound and Language Processing' of the Greek General Secretariat for Research & Technology (GSRT).

6. REFERENCES

[1] M. J. Black and A. D. Jepson. Recognizing temporal trajectories using the condensation algorithm. In *3rd. international Conference on Face and Gesture Recognition*, 1998.

[2] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic Active Contours. *Int'l J. Computer Vision*, 22(1):61–79, 1997.

[3] E. Efthimiou and S.-E. Fotinea. GSLC: Creation and annotation of a greek sign language corpus for hci. *Universal Access in Human Computer Interaction. Coping with Diversity*, pages 657–666, 2007.

[4] G. Fang, W. Gao, and D. Zhao. Large vocabulary sign language recognition based on fuzzy decision trees. *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, 34(3):305–314, 2004.

[5] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of Imaging understanding workshop*, pages 121–130, 1981.

[6] V.-M. Mantyla, J. Mantyjarvi, T. Seppanen, and E. Tuulari. Hand gesture recognition of a mobile device user. In *IEEE International Conference on Multimedia and Expo*, 2000.

[7] S. Ong and S. Ranganath. Automatic sign language analysis: a survey and the future beyond lexical meaning. *Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891, 2005.

[8] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed:algorithms based on Hamilton-Jacobi formulations. *Journ. of Comp. Physics*, 79:12–49, 1988.

[9] I. Ozer, L. Tiehan, and W. Wolf. Design of a real-time gesture recognition system: High performance through algorithms and software. *Signal Processing Magazine, IEEE*, 22:57–64, 2005.

[10] N. Paragios and R. Deriche. Geodesic active regions: A new framework to deal with frame partition problems in computer vision. *Journal of Visual Communication and Image Representation*, 13(1/2):249–268, March 2002.

[11] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[12] H. Wang, M. C. Leu, and C. Oz. American sign language recognition using multi-dimensional hidden markov models. *Journal of Information Science and Engineering*, 22, 5:1109–1123, 2006.

[13] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(9), 1999.

[14] S. Wong and R. Cipolla. Continuous gesture recognition using a sparse bayesian classifier. In *18th international Conference on Pattern Recognition, ICPR. IEEE Computer Society*, 2006.

[15] Y. Wu and T. Huang. Hand modeling, analysis, and recognition for vision-based human computer interaction. *IEEE Signal Processing Magazine*, 18:51–60, 2001.

[16] M.-H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(8):1061–1074, 2002.