HMM-Based Continuous Sign Language Recognition using a Fast Optical Flow Parameterization of Visual Information

G. Cortés, L. García, C. Benítez and J.C. Segura

Dept. Signal Theory, Networking and Communications
University of Granada, Spain
gcortes@correo.ugr.es, luzgm@ugr.es, carmen@ugr.es, segura@ugr.es

ABSTRACT

This paper presents a preliminary study of an optical flow-based parameterization of visual information in a sign language recognition system using Hidden Markov Models (HMM). Current feature extraction processes need initialization, tracking and segmentation stages in order to describe signer gestures. Our aim is to develop a single and fast technique to reduce computational complexity which doesn't require these stages and is able to work in mobile devices with limited hardware resources. The *Moving Block Distance (MBD)* parameterization is an interesting first approach for this purpose, proved by two signers under a static background constraint. A lexicon of 33 basic word units *(signemes)* was used to build the data set containing phrases with a variable number of words. Continuous recognition results achieve more than 99% accuracy in close test.
Index Terms: multi-modal recognition, visual feature extraction, mobile devices, sign language, optical flow.

1. INTRODUCTION

The communication among different people is one of the major challenges of the mankind. Deaf people community is not an exception and the sign language is its main way of communication. Usually (like in spoken languages) each country has its own sign language even with dialects. Current machine translation of sign language demands language scalability, portability to mobile environments, real time and continuous recognition functionalities at the same time in order to work properly. The first step to implement machine recognition is to learn how communication occurs between sign language speakers (signers). In a conversation, signers do a multi-modal analysis from hands and facial gestures although hands are enough for a first approach. Machines usually need tracking and segmentation methods to isolate these information sources. Recent approaches need specialized hardware (like 3D trackers and datagloves) and powerful computers to extract the set of features and recognize in real time. Features observed by signers (perceptual parameters) are divided in manual (handshape, handorientation, location and motion) and non manual (gaze, facial expression, mouth movements, position and motion of the trunk and head) [1].

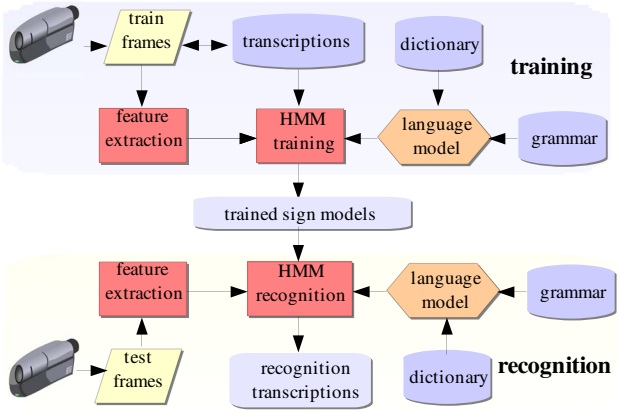

Figure 1: Structure of the HMM-based visual recognition system.

Machine recognition systems normally perform their own parameterization (feature vectors) emulating manual features.

Due to the multi-modal character of the sign language many of its issues could be solved using methods of spoken language and computer vision areas. Typical problems involved in the recognition of visual gestures are [1]:

- Hand and face occlusions.
- Temporal / spatial sign boundaries.
- Segmentation and tracking of hands and face.
- Coarticulation: signs are affected by their neighbors.
- Variation of the signer position / posture.
- Variance (in time and 3D space) of the signs.
- 3D scene projection in a 2D image plane produces information loss while parameter extraction in 3D scenarios is a time consuming process and can risk real-time functionality.

HMM have the ability to fix most of the problems given in speech like word delimitation, time variances, real-time operation, coarticulation and vocabulary scalability. In fact, is the main technique in sign language recognition [1-5]. HMM apply a statistical approach of finite state machines useful on pattern learning and recognition [6]. Computer vision troubles as segmentation, tracking, 3D reconstruction, occlusions and classification are handle by miscellaneous approaches based on motion estimation, color information, edge detection, fuzzy logic, artificial neuronal networks or statistical methods [7] .

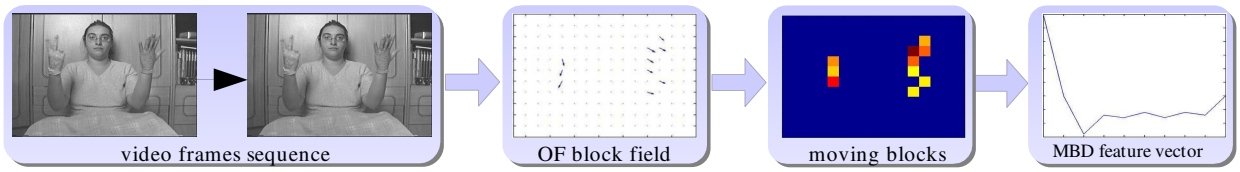

Figure 2: Feature extraction technique. The moving blocks between two video frames are computed using optical flow (OF) and the relative distance between them (MBD) is calculated to achieve one feature vector per frame transition.

2. AUTOMATIC MACHINE SIGN LANGUAGE RECOGNITION

An automatic recognition system for sign language has the same structure and stages that a speech recognition system but describes visual features instead of sound-based features (Fig. 1).

2.1 Sign Language Approaches

Visual data acquisition stage defines two main sign language recognition approaches [3]: dataglove and video-based systems. Dataglove systems have embedded sensors in gloves usually connected to hardware which extracts parameters of the motion, trajectory and shape of the hands. This allows real time 3D feature extraction without initialization and segmentation stages but it is uncomfortable, expensive and non-portable to real scenarios. In video-based systems a set of video cameras records a sequence of images to process with certain software. It's a more natural human-machine interface but needs more environment constraints (background and clothes colors, signer posture and location, etc.) in order to extract parameters from the images. With more than one camera 3D features can be extracted but harder processing is required. Normally initialization, tracking and segmentation stages are required and similar results than dataglove systems are achieved. Hybrid methods can be used to improve recognition accuracy in mobile scenarios [5].

2.2 Object description: visual features

Several sets of features are used to describe visual signs. Feature selection is oriented to obtain the greatest reliability/complexity ratio to ensure signer independent recognition and real time operation. We can find two basic groups of features:

i) *Geometric features of objects*: sizes, areas, contours, relative and global positions, velocities, accelerations and angles are some examples [4].

ii) *1D projection of 2D edges:* curvature function (measurement of the changes in the curvature of an object contour) and distance function (between n-relevant points and one reference point).

Data extraction process in video-based systems is the same as in computer vision and requires segmentation, tracking and description of the objects. Objects correspond to hands in first approximation and head, arms and trunk for a deeper analysis. Segmentation is often used as initialization of tracking algorithms [4]. Localization templates allow to locate our objects (hands, arms, elbows, clothes and head) by their position and later extract some features to track them. Features often used to segment are object textures, colors and shapes. Color segmentation is the main way of object tracking. For this purpose, to wear colored clothes and gloves is a useful constraint. Spatial prediction, velocity and acceleration of the already segmented objects are used to track, to solve occlusion problems and to reduce computational complexity [4]. Classic computer vision methods like Kalman's filter, particle filter and active shape models can work fine to segment and track objects [7].

3. OPTICAL FLOW AS VISUAL GESTURE FEATURE: MOVING BLOCK DISTANCE

Alternatively, we can use the optical flow (OF) to build the feature vector. OF is the apparent motion of brightness patterns (i.e., the mentioned *objects*) of an image sequence [8] and is represented by 2D motion vectors forming the motion field. At first sight OF has interesting properties for our purpose:

- Many and efficient methods to obtain OF vectors are available. Some of them allow real time operation with low computational cost.
- Good characterization of the parameters used in the recognition by signers: both, perceptual manual and non-manual parameters can be described. Furthermore, OF can model face gestures and head and trunk motion so is no limited to hand description.
- Object segmentation is not required: with a static background the motion of the significant objects (hands, arms and maybe head and trunk) is well described by the motion vectors with greatest module (see Fig.3). Indeed, OF is often used to segment objects in motion.

Typical problems of OF estimation like hand occlusion and aperture can appear. These matters can be solved selecting appropriate colors and textures for clothes, wearing colored gloves and using color segmentation. However, we analyzed our video frames in gray color scale and without any segmentation. Other backwards must be considered:

- Shadows of signer's hands/head can be a problem if they don't appear close (in the same block) to the hand edges.
- Objects like glasses, hair, hats..., signer body shape and 3D relative position between signers and camera can negatively affect the motion field adding signer dependent components.

In order to achieve signer independent and a small visual set of features we'll use a single parameterized form of motion vectors: the Moving Block Distance (MBD). A frame is divided in blocks. The blocks are numbered from left to right and from

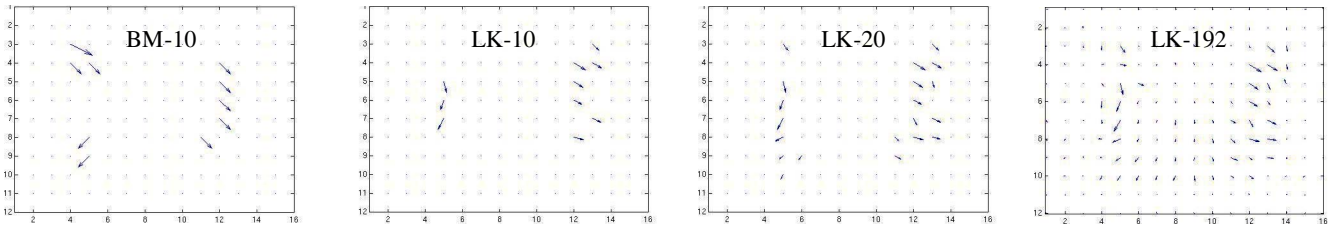

Figure 3: Block Matching (BM) versus Lukas-Kanade (LK) Optical Flow (OF) computation methods. We compare BM with 10 vectors and LK using 10, 20 and 192 (full block motion field) major OF vectors.

top to bottom of the frame so (x, y) block positions are indexed into a mono dimensional value *(p)*. A motion vector is calculated for each block building the vector motion field. The block motion field can be well characterized only with the 10 major motion vectors, naming the so-called *moving blocks* for each frame (see Fig. 2). The difference between the (p) values of consecutive (from left to right and from top to bottom) moving blocks is then calculated and stored forming the parameter vector for each frame. We define this difference as Moving Block Distance.

Two issues must be taken into account in the MBD parameterization. The first is related to the block size: large block sizes achieve low bitrates and signer independent description of the hands, head and trunk. Dense block motion field and little block sizes mean too many signer dependent and scenario details and higher spatial resolution. Finally, a concealed size of 20*20 pixels was selected. The second issue refers to the OF computation. We analysed two simple methods for this purpose: Lukas-Kanade (LK) and Block Matching (BM). LK seems to obtain a slightly better description of signer motion so was the selected method for our experiment.

A spectral analysis (Fig. 4) reveals that different visual words have slightly different spectrums and how the same word maintains its spectral shape even when is signed by two persons. This fact confirms that MBD technique can be used to perform independent signer recognition and hopefully, recognition of non previously trained data (blind tests).

4. EXPERIMENTATION

4.1 Data acquisition

We built three data sets with two non professional signers. Each signer recorded the data of one set and later these two sets were merged to create a third one. We used a single analog color camera recording 25 frames/sec of 320*240 pixels. All the videos were digitalized and the frames extracted in gray color with 8bpp. 31 videos for the signer 1 and 28 for the signer 2 were recorded with a variable number of visual sentences.

4.2 Language modeling

We adapted the "voice-dialing" example of the HTK manual [9] to the Spanish sign language. A total of 33 visual signs were used including the *pausa* model corresponding to a rest/initial pose plus the *start_end* and *sp* models used as sentence/sign delimiters (Table 1). Proper nouns (as *Robert Smith*) were decomposed in

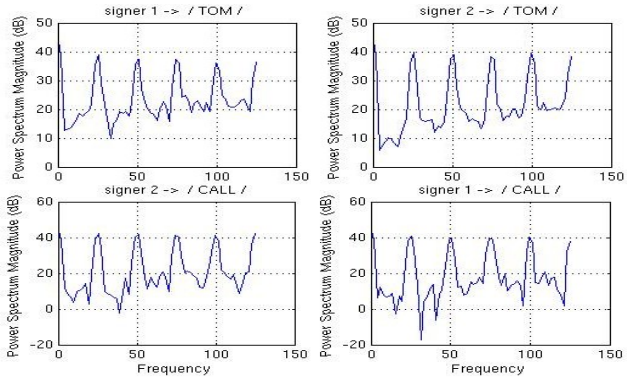

Figure 4: Power Spectral Density of parametrized visual words. The same words signed by two persons have similar spectrum patterns while different words have a slightly different shape.

visual alphabet letters (signeme level). The signer data set contains 54 visual sentences and 20 minutes of total recording time. A grammar constraint to build sentences with some sense was used with no fixed length (from 3 to 20 words) and no fixed number of sentences (from 1 to 4) in each video.

Table 1: Lexicon used and sign decomposition. One HMM per signeme (basic sign unit) was used.

	visual words	*signemes*
verbs	TELEFONEA, GRABA, MARCA	telefonea, graba, marca
numbers	CERO, UNO, DOS, TRES, CUATRO, CINCO, SEIS, SIETE, OCHO, NUEVE	cero, uno, dos, tres, cuatro, cinco, seis, siete, ocho, nueve
proper nouns	TOM, BAILEY, ROBERT, SMITH, CAETANO, VELOSO, DAVID, BOWIE, BYRNE	t o m, b a i l e y, r o b e r t, s m i t h, c a e t a n o, v e l o s o, d a v i d, b o w i e, b y r n e
delimiters	START_END, PAUSA, SP	start_end, pausa, sp

4.3 Parameterization, training and test stages

The feature vector stores 10 MBD samples for each frame with 2 bytes of precision and 25 fps achieving a low bitrate of 0.5 KB/s and 250 samples/s. No overlapping features and no temporal smooth was performed. Our system is based on Continuous Hidden Markov Models (CHMM) with simple left to right topology and 15 emitting states. New components were

incrementally added to the output probabilities in each training stage, from 1 to 5 and finally to 25 multivariate Gaussians Probability Density Functions (PDFs). We used an autoreestimation technique to control the number of training reestimations of the model set *M* using as flag the normalized increment in the probability of *P(O|M)* where *O* is the sequence of observable events parameterized by the feature vectors.

Three close tests were built, one for each data set. The same data and configurations were used for training and testing. Table 2 shows recognition results.

Table 2: Close test recognition results.

signer	*HMM-reests*	*Acc%*
1	9 -> 32 -> 47	59.17 -> 83.15 -> 99.54
2	10 -> 28 -> 49	62.54 -> 95.13 -> 99.71
1+2	10 -> 27 -> 49	62.54 -> 88.21 -> 99.25

Table 3. Related work comparison with blind ([1, 2]) and close ([3, 4, 5, this]) tests.

work	*Acc %*	*Signs*	*Train set words/time*	*features rate*
[1]	95	52	? / 210min	30*13fps
[2]	91,7	97	? / 210min	30*13fps
[3]	99.4	40	2500 / ?	16*10fps
[4]	93	70	?	14/frame
[5]	98.05	141	665*4 / ?	>22/frame
[this]	99.50	33	906 / 40min	10*25fps

5. RESULT DISCUSSION AND CONCLUSION

In Table 3 we take a look at the state of the art. Only Tanibata *et al.* [4] do not use language model constraints. Bauer *et al.* [1-2] perform a blind test, so their results in close test probably reach 99%. All of them (but this paper *[this]*, in the mixed test) do a recognition based on a test data build by an unique signer. The MBD parameterization of motion estimation (and this study) is focused to obtain a fast feature extraction process with the lowest computational cost. Our result (Table 2) corresponds to the average of the three test results and despite of using few train data proves that MBD is an interesting approach. The final challenge is to embed the whole recognition system as a software package in a mobile device like a cell phone or PDA which works in real scenarios. Future improvements of MBD technique will be focused to achieve signer independent recognition and to solve the static background constraint. Development process would include:

- Study of faster and stable methods of motion estimation.
- To increase the number of moving blocks per frame and test the accuracy rate.
- Background noise models can be trained to raise robustness in mobile scenarios.
- A greater database would lead to independent signer and blind test recognition and robust models.
- Spatial normalization: a dynamic zoom into the interest region of the scenario. The 3D information loss could be compensated by this normalization.
- The addition of differential MBD parameters to the feature vector could reduce signer dependent problems (like glasses, hair, hats, and so on).
- Interoperability with many kind of mobile devices of diverse camera resolution and frame rate could be performed 'on the fly' changing the block size but keeping their number. Pyramidal methods of interpolation can be useful.
- The number of logic-states of a HMM must be proportional to the time duration of its related signeme. This would increase the robustness of the system.

6. REFERENCES

[1] B. Bauer, H. Hienz, K-F. Kraiss. “HMM-based continuous sign language recognition using stochastic grammars”, *Proc. of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pp. 185 - 196. 1999.

[2] B. Bauer, H. Hienz. “Relevant features for video-based continuous sign language recognition”, *Proc. of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition.,* pp. 440-445, 2000.

[3] T. Starner, J. Weaver, A. Pentland. “A wearable computer based american sign language recognizer”, *Proc. of the 1st IEEE International Symposium on Wearable Computers*, pp. 130-137, 1997.

[4] N. Tanibata, N. Shimada, Y. Shirai “Extraction of hand features for recognition of sign language words”, *Proc. of International Conference on Vision Interface*, pp.391-398, 2002.

[5] R.M. McGuire, J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear, D.S. Ross “Towards a One-Way American Sign Language Translator”, *Proc. of the 6th IEEE International Conference on Automatic Face And Gesture Recognition*, pp.620-625, 2004.

[6] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, Vol. 77, No. 2, Feb 1989.

[7] A.M. Tekalp, *Digital Video Processing*, Ed. Prentice-Hall, 1996.

[8] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. *Performance of optical flow techniques,* International Journal of Computer Vision, Vol. 12, No. 1, pp 43-77, 1994.

[9] S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev. And P. Woodland, *The HTK Book (Version 3.2)*, Cambridge University, 2002.