Body posture estimation in a sign language video

F. Lefebvre-Albaret and P. Dalle

IRIT : UPS-118 r. de Narbonne, 31062 Toulouse cedex 9

Key words: Sign Language, Posture Reconstruction, Inverse Kinematics

In the framework of depth estimation from mono camera videos, several techniques such as shape-from-shading, shape-from focus/defocus, shape-from-motion and shape-from-texture can be combined to determine the depth in each frame. They are based on hypothesis which are not confirmed in sign language gestures. The results are improved if a geometrical or dynamical model is known before the depth estimation. In the frame of body posture estimation, we can then apply inverse kinematics algorithms. The use of SL features can also enhance body posture estimation. In this paper, we present an algorithm that takes both geometrical model and SL specificities into account to afford a 3D reconstruction of arm posture that could be used in SL processing.

The analysis of mono-camera SL videos involves several difficulties: camera calibration, variation of signer appearance, frequent occultation of the body parts (like hands, head and shoulders), difficulty to localize body joints accurately because of clothes, correspondence of at most four possible arm postures (even if the 2D position of hand, elbow and shoulder have been precisely determined), number of degrees of freedom of upper body part.

Since the hands are often in front of the torso, occluding more than the half of each arm, we must combine several approaches for the body part tracking, as in [1] . We have chosen to use both silhouette and colour features. SL features are used to disambiguate hands, to track the elbows and to estimate the posture more accurately. Taking SL features into account in each steps of the tracking makes our algorithm both original an well adapted to continuous SL processing.

Our method begins with the location of the following body parts:

The Head and the two hands are tracked thanks to their colour. The skin model is learned from several skin pixels through skin images and their detection is then based on a bayesian approach.

The elbows are tracked from the silhouette thanks to their appearance. As most signs of sign language are executed in the neutral space (space in front of the signer), the elbows can be detected as being corners in the silhouette. This criterion sometimes give outliers that can be removed in limiting the displacement of each elbow between two consecutive frames.

A step of disambiguation is necessary to label the right and the left hand correctly. We designed an original algorithm based on dynamic programming taking into account the hand relative position, the concept of dominating hand (specificity of sign language), the distance to the elbows and the motion continuity to solve this problem. The results of the disambiguation are 98,7% (the

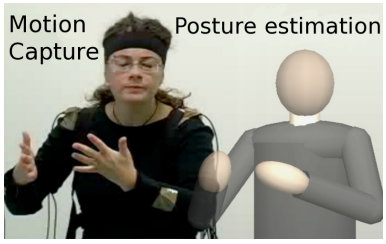
two hands where crossed 10,1% of the time in our test sign language videos).
The tracking of the shoulders is not obvious because of their frequent occlusion by hands. These occlusions result from an intensive use of the space around the face in sign language. As a consequence, their position have been deduced from the head position (Our experiment showed that the 2D distance on x and y axes between head and shoulders is quite stable in each frame).

What must be the precision of the depth estimation so that the posture reconstruction can be use in SL analysis? To answer this question, we must consider the role of the hand depth in sign language. Depth is used to **instantiate or to refer to entities** in the signing space. In this case, the absolute hand position has to be respected. Depth is also involved in **directional signs** such as "I give you" (relative motion from the previous frame is important). In the **bi-manual signs**, hands must be placed in an accurate relative position. Depth estimation must reach a compromise between those three criteria.

To effectuate the posture reconstruction, we use the approach described in [2]. The posture is estimated from the hands, elbows and shoulders positions. By now, the rotation movement of the torso is ignored because its contribution to the hand depth is small compared to those of the arm and the forearm. However, it will have to be taken into account in a further version. The differentiation of the equation used for the depth estimation shows that a small error on body part position estimation can lead to important inaccuracy on the depth estimation. This fact explains why the noise of row measures do not allow their direct use.

To remove the noise from the measure, we first used sign language specificities. When we compare right and left hand depths, we measure a correlation of 0,24. The relationship between the two hands motion is used and combined with a Kalman smoother to reduce the noise and make the measure usable both in the generation by a virtual signer and in a context of an automatic processing.

We evaluated our algorithm on a three-minute long video. We observed an improvement of the correlation between the ground truth depth estimation of hands position and the calculated depth. The correlation is 20% better when we take the correlation between the two hands into account. The smoothing of the movement makes it more natural and the relative position of the two-hands is better respected during the two handed signs. Other tests must be lead to validate our approach on bigger corpora.



References

1. Leignel, C., Viallet, J.E., A blackboard architecture for the detection and tracking of a person, RFIA2004, Toulouse (France)
2. Lenseigne, B., Gianni, F., Dalle, P., A new gesture representation for Sign Language analysis, LREC 2004, Lisbon (Portugal), p84-90