Head Tracking and Hand Segmentation during Hand over Face Occlusion in Sign Language

Matilde Gonzalez, Christophe Collet, and Rémi Dubot

IRIT (UPS - CNRS UMR 5505) Université Paul Sabatier,
118 Route de Narbonne,
F-31062 Toulouse Cedex 9, France
{gonzalez,collet,dubot}@irit.fr
http://www.irit.fr

Abstract. This paper presents a method to accurately segment the hand over the face. The similarity of colours and the important variability of the hand shape make it challenging. We propose a method based on the combination of two features: pixel colour and edges orientation. First, a specific skin model is used to find, before occlusion, the face position and the face template. Then, during occlusion the face template is registered using local gradient orientations to track the face position. Colour information is extracted from changes on pixel colours and edges are classified as belonging to the hand or to the face by mapping edges orientation to the face template. Finally by merging both features and by using an hysteresis threshold, which considers connectivity, a robust hand segmentation is reached. Experiments were performed using the Dicta-Sign corpus and showed the versatility of the proposed approach.

Key words: Hand segmentation, sign language, head registration

1 Introduction

Deaf and hearing-impaired communities use sign language to communicate. It is a visual language characterized by the motion of the mouth, eyes, face, trunk and hands. Nowadays many researches focus on the automatic analysis and recognition of sign language, especially, automatic sign language interpretation [1–3]. This would enable sign language users to communicate to anyone without the need of human interpreter. To achieve this, linguistic models are built and video treatments developed. The annotation of sign language corpus plays an important role for doing some statistics needed for linguistic models or for evaluating automatic treatments. The annotation is, in general, manually performed by linguists and computer scientists through several annotation tools, e.g. Elan [4], Anvil [5], Ilex [6, 7], Ancolin [8], etc. For long video sequences, manual annotation becomes error prone, unreproducible and time-consuming. We have proposed [9] a distributed system architecture to assist the annotation using automatic video treatments trough the network. In this paper we introduce a video treatment to assist the annotation of hand shape and position. Since the face and hands often

overlap and because these regions are similarly coloured, hand shape recognition becomes really challenging. This method will be used with the tracker introduced by Gianni *et al.* [10] which will be modified to detect occlusions in order to find the template frame before occlusion.

In previous works, hand region is obtained by assuming the hand to be the only object in the image [11] or by extracting skin colour regions [3, 12]. However, these approaches do not handle skin objects occlusions. Others approaches based on active contours [13, 14] give good results but do not cover the fast change and variability of hand shape. They assume that hand shape change is very small between successive frames which is, normally, not the case in sign language unless the video is acquired on specific recording conditions such as high-speed frame recording. In [15] is introduced a template based approach. They consider the face and hand template before occlusion. Even though face deformation remains small, 2D hand shape, during occlusion, can quickly change without any hand configuration changing. In [16] an approach to solve hand over face occlusion is introduced using the concept of image force field. The results showed that the hand is roughly segmented. This might not be enough to find hand configuration for sign language analysis.

In this work, we focused on face tracking and hand segmentation during hand over head occlusions. Our approach is not hand model based because hand changes very fast. However, it uses a before occlusion template of the face. The main idea is to find any information that was not present before the occlusion. We assume, then, that any change from the template is caused by the hand. We noticed that only considering colour feature was not enough because of the huge similarity between face and hand. That is why we decided to use edges information to complement colour features. The main advantage of our approach is that it is not dependent on 2D/3D hand shape from previous frames. Our work has been tested in the LSF Dicta-Sign corpora [17], in which each video is mono-camera and contains one signer recorded in a frontal viewpoint and an homogeneous background.

In the section to follow the skin model initialisation and skin colour modelling distribution are described. Section 3 explains the head position initialisation from the skin segmentation and the head registration using a template to track the face in occluded images. Section 4 details how the hand region is extracted from the occluded image by merging edges and colour features. Finally, section 5 shows the qualitatively and quantitatively evaluation performed to point out the quality of the results and the limitations of the approach.

2 Skin Segmentation

Sign language features have a common skin colour that depends on the signer: hands and face. Many approaches have been developed for skin segmentation [18]. Some of them need a training step and make the model dependent to the skin-colour samples, illumination and environment conditions in the training set. Using a specific model, i.e. a model built using skin pixels from the same im-

(a) (b) (c) (d)

Fig. 1: (a) shows the image without occlusion, (b) is the segmentation result from Eq. (1). Notice that the neck region is no detected. (c) is the connected region used as sample skin pixels and (d) show the final segmentation result.

age/subject to treat, makes the model more specific and the segmentation more robust. It is necessary to first find skin regions to train the specific model. It is automatically done using a general skin model. Thus the model is specific to the subject and to the recording environment.

2.1 Sample Pixels Initialisation

In this work we use explicitly defined skin regions for initialisation. Kovac *et al* [19] defined, through a number of rules, the boundaries skin cluster in RGB colorspace:

$$\begin{aligned} &\textit{(R,G,B) is classified as skin if:} \\ &R < 95 \textit{ and } G > 40 \textit{ and } B > 20 \textit{ and} \\ &max\{RGB\} - min\{RGB\} > 15 \\ &|R - G| > 15 \textit{ and } R > G \textit{ and } R > B. \end{aligned} \tag{1}$$

The main difficulty achieving high recognition rates with this method is that the luminance component and the chrominance components are not decoupled. For example using Eq. 1 for each pixel of Fig. 1a, the result, Fig. 1b, shows that in regions where a shadow appeared the detection became inaccurate, e.g. neck and fingers. However the advantage of this approach is that there is no need of any learning stage, it is easily implemented and it gives a rough skin sample region. The sample pixels for the model are those belonging to the greatest area connected component, head or hand(s) depending on the frame, as shown in Fig. 1c.

2.2 Specific Model Generation

The skin-colour is modelled as a bivariate normal distribution in the YCrCb colorspace. The Y component reflects the luminance and is rejected to solve

(a) (b) (c) (d)

Fig. 2: Head registration results for a sequence of frames. (a) is used as a template and is registered to the following frames in the sequence, e.g. (b),(c) and (d).

shadow problems. The mean vector μ_S and the covariance matrix $\mathbf{\Sigma}_S$ of the distribution are estimated from the skin training pixels and used to compute the threshold values for segmenting any following frame. Thus the skin classification is a binary decision and the cut-off values are then automatically computed for each signer. Fig. 1d shows the segmentation result. We notice that the skin pixels are better detected. However, because of the simplicity of the decision rule, some pixels belonging to the hair were wrongly assigned to the skin class. We do not expect this to pose a significant problem since in this work our interests lie in the segmentation of the hand including edges information.

3 Head Tracking

3.1 Head Position Initialisation

Face position is determined using the skin segmentation explained in the previous section. For the first frame of the sequence (before occlusion), we consider that the head is the connected region with the greatest area placed in the upper part of the frame, Fig. 2a. Texture and edge orientation of the face region are registered as templates. In the future we will consider that the head position is given by the tracker [10].

3.2 Head Registration

When the face is occluded by the hand it is not possible to accurately find the head position using skin blobs. It is necessary to use another property than colour. We propose to extract edges information from the face template to find the head in an occluded image. The template is aligned using the direction of the gradient after a threshold. The image transformation matrix considers displacements in $\mathbb{R}^2$ and rotation in the plane $X \perp Y$. Out-of-plane rotations are not handled in this paper. The optimisation function is defined as

$$d(x, y, \theta) = \arg \min_{\theta \epsilon [\theta_{min}, \theta_{max}]} \sum_{(x,y) \epsilon I} \min({\theta_T}'(x + x', y + y') - \theta_I(x, y)), \quad (2)$$

where $(x', y')\epsilon N \times N$ with N the size of the searching window, θ_{min} and θ_{max} are, respectively, the minimal and maximal rotation angles, θ'_T represents the edge orientation of the image template rotated by the angle θ. For instance the optimisation search is performed in a large neighbourhood inside a window where the size is large enough to allow substantial head movements. Even though this searching algorithm is not optimized, local minima are avoided. The alignment of the template face handles local face deformation (e.g. lips, eyes, etc.) and partial occlusions leading to good results as long as the out-of-plane face rotation remains small. Fig. 2 show some results of the head registration. We intend to use this registration coupled with the tracker [10] to improve results during occlusions.

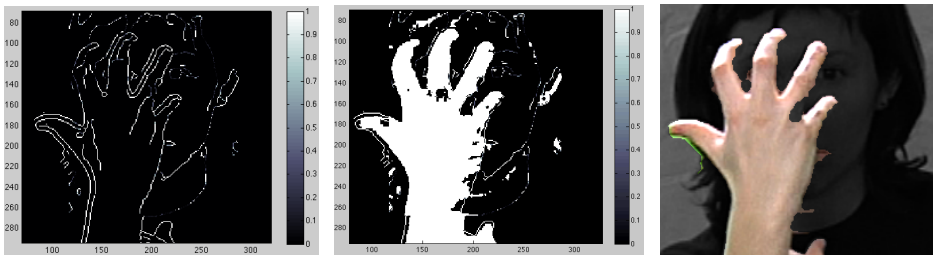
4 Hand Extraction

The segmentation of the hand is not easily performed by only considering colour feature. Edges information has to be used to well define hand boundaries. That is why we intend to classify edges into two classes: edges belonging to the hand and those belonging to the head. We first compute a pixel-to-pixel edge orientation difference map. Then we use the colour information from pixels that have considerably changed to determine hand pixels. Finally using pixels connectivity we are able to extract the hand.

4.1 Edge Orientation Difference Map

We use Canny edge detector in both the template image and the occluded image, with a defined threshold, to detect edges. For each pixel edge that belongs to the occluded image we search the closest edge, within a defined neighbourhood, in the template. When an edge is found we compute the orientation difference which is the minimal angle between the two directions. Otherwise when no edge has been found in the defined neighbourhood the difference is considered the highest orientation difference value ($\pi/2$). That means that this pixel in the occluded image has an large probability to belong to the hand. The orientation difference map is built in this way for each edge pixel in the image with occlusion. Fig. 3a shows the normalized orientation difference map of the image in Fig. 2d. Notice that most of the edges belonging to the hand have high values, close to ($\pi/2$), and many edge pixels from the face have low values, close to 0. In places where edges from the hand intersect edges from the face, e.g. over the mouth or eyes, other values appear depending on the intersection angle (low values if edges coincide).

The separation of the edges is not straight forward from this difference map. For low values it is not easy to define if edges coincide or if they really belong to the face. Then if we try to classify edges with no further information we can miss important information. That is why the next step is to use pixels colour changing to remove this ambiguity.



(a) (b) (c)

Fig. 3: (a) shows the normalized edge orientation difference map. (b) shows pixels that have changed and edges orientation difference map merged. White pixels have greater probability to belong to the hand. (c) shows the hand segmentation result, dark gray and black pixels belong to the non hand class.

4.2 Colour and Edges Combination

Colour feature is used to complement information. We noticed that during occlusions, in places where there is an ambiguity concerning edge classification pixels colour have considerably changed. For example in the mouth or eyes area hand colour is very different, however edges could coincide depending on the shape and position of the hand over the face. On the other hand in some other places, like the cheek or the forehead, the colour is very similar but edges are easily distinguished. These examples shows why merging these features makes the classification more robust.

Colour contribution is computed by a simple pixel-to-pixel colour difference. A threshold is applied to obtain the major colour changes. We have added it to the difference map built before Fig. 3b. We apply the Hysteresis threshold, which considers connectivity, to segment the hand. We have filled the holes and extracted the largest connected area (hand). The segmentation result is shown in Fig. 3c.

5 Evaluation and Results

This section presents the results and evaluation of the hand over face segmentation methodology. Because of the lack of publicly-available annotated corpus we have used several frame sequences from the LSF Dicta-Sign corpus where the pixels belonging to the hand have been manually annotated. The corpus is composed of several sign language conversations between two signers with a total of 16 different signers and a total of 5 hours of video. We have selected some sequences that contain hand over face occlusion. In these sequences, hand shape and face expression can change during the occlusion. Our approach has been tested on 5 sequences, around 50 images with face occlusion. In each sequence the first frame is assumed to be without occlusion and is used as template. All the following frames contained hand over face occlusion.

Fig. 4: First row shows five consecutive frames in a sequence with hand over face occlusion. The second row shows the segmentation result. Pixels in dark gray or in black have been classified as belonging to the non hand class, otherwise they are shown in their natural colour.

Fig 4 shows in the top row the images to segment and in the bottom row the segmentation results. Notice that in the segmentation results, pixels that have been classified in the hand class are shown in their natural colour, otherwise they are shown in dark gray or in black. Fig 5 shows the segmentation results for several frames of various sequences. These two figures indicate that the hand over face segmentation was performed reasonably well, however we can see some artefacts and some holes. The artefacts are mainly due to large pixel changes, e.g. out-of-plane rotation or/and substantial face expression changing, or wrong skin segmentation. On the other hand the holes are caused by a lack of information. In fact sometimes an edge from the hand can coincide to an edge from the face, in terms of orientation and position. In that case the hand edge might be classified as belonging to the face. Then if there is no colour information because the pixel colour remains very similar, some pixels will be wrongly classified. In any case the overall hand shape is well defined.

The performance of the proposed segmentation approach has been qualitatively evaluated. Now to quantitatively evaluate this method, we manually generated ground truth segmentation for all the frames in the sequences. Since the evaluation performed is pixel wise, the ground truth is a binary image of hand pixels. These ground truth images are used as reference to compare the automatically segmented images. The true positive (TP) and the false positive (FP) percentages are evaluated for each image of the sequences by

$$TP(\%) = \frac{Number\ of\ correctly\ detected\ pixels}{Number\ total\ of\ hand\ pixels} \times 100 \tag{3}$$

$$FP(\%) = \frac{Number\ of\ wrongly\ detected\ pixels}{Number\ total\ of\ hand\ pixels} \times 100 \tag{4}$$

where TP(%) corresponds to the rate of correctly detected pixels with respect to the total number of pixels to be detected and FP(%) to the wrongly detected

Fig. 5: This figure shows the segmentation results for 3 different sequences. Each row corresponds to a sequence.

Fig. 6: Segmentation results: artefacts under the chin and over the collar

pixels with respect to the total number of pixels that should not be detected. Since FP(%) depends on the number of non hand pixels, this rate becomes dependent of the background and size of the image. For this reason we decided to compute the FP(%) rate with respect to the total number of hand pixels. Thus this rate is representative to the number of hand pixels on the image. Table 1 presents the rates evaluated for each sequence, we notice that the TP(%) rate is in average about 96%, reaching until 99% for some frames. The FP(%) rate is about 8% and corresponds to pixels that can be easily detected and eliminated by post-treatments, e.g. thin lines under the chin and/or over the collar in Fig. 6. Moreover we plan to extract global features (e.g. number of fingers, global shape, etc.) to characterize the hand and these measurements should not be corrupted by the remaining artefacts.

6 Conclusion and Future Work

In this paper we have detailed a hand over face segmentation approach. It takes into account two local features; colour and edges. It considers the pixels colour

Table 1: Results of the evaluation rates trough several sequences

RATES	SEQUENCE #					AVERAGE
	2	4	7	9	11	
TP(%)	96.71	96.51	96.59	95.07	98.15	96.61
FP(%)	3.62	6.48	13.91	6.44	8.27	7.74

changing and the edges orientation. When the hand occludes the face, both features information complements each other. That means that in some places where the pixels colour remains quite similar, we still have the edges information and vice-versa. By merging these features and by comparing the occluded image to the face template, we are able to well segment the hand. Experimental results indicated that this method is able to extract the hand effectively and showed the limits of the approach regarding the quality of the segmentation: artefacts and holes. The constraints imposed by treating sign language video corpora allow us to validate this algorithm and to focus on the analysis of hand configuration in sign language. It is important to be aware of these limitations if the same methodology is to be used for other purposes.

In the future, we intend to improve this segmentation by solving the out-of-plane rotation, improving the skin detection and edges rejection rules. We also envisage to use this segmentation algorithm in tracking systems to accurately detect the position of the hand when it occludes the face which is one of first limitations in tracking systems.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n231135.

References

1. Imagawa, I., Matsuo, H., Taniguchi, R., Arita, D., Lu, S., Igi, S.: Recognition of local features for camera-based sign language recognition system. In: Proc. 15th International Conference on Pattern Recognition. Volume 4. (2000) 849–853
2. Liang, R., Ouhyoung, M.: A real-time continuous gesture recognition system for sign language. In: Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings. (1998) 558–567
3. Habili, N., Lim, C., Moini, A.: Segmentation of the face and hands in sign language video sequences using color and motion cues. IEEE Transactions on Circuits and Systems for Video Technology **14** (2004) 1086–1097
4. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: Proc. of the 5^{th} International Conference on Language Resources and Evaluation (LREC 2006). (2006) 1556–1559
5. Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue. In: Proc. of 7^{th} European Conference on Speech Communication and Technology (Eurospeech). (2001) 1367–1370

6. Hanke, T.: ilex - a tool for sign language lexicography and corpus analysis. In: Proc. of 3^{rd} International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas de Gran Canaria, Spain (2002) 923–926
7. Hanke, T., Storz, J.: ilex - a database tool for integrating sign language corpus linguistics and sign language lexicography. In: Proc. of 6^{th} International Conference on Language Resources and Evaluation, LREC 2008, Marrakesh (2008) W25–64–W25–67
8. Braffort, A., Choisier, A., Collet, C., Dalle, P., Gianni, F., Lenseigne, B., Segouat, J.: Toward an annotation software for video of sign language, including image processing tools and signing space modelling. In: Proc. of 4^{th} International Conference on Language Resources and Evaluation - LREC 2004. Volume 1., Lisbon, Portugal (2004) 201–203
9. Christophe Collet, Matilde Gonzalez, F.M.: Distributed system architecture for assisted annotation of video corpora. International workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC), Valletta, Malte (2010) 49–52
10. Gianni, F., Collet, C., Dalle, P.: Robust tracking for processing of videos of communications gestures. Gesture-Based Human-Computer Interaction and Simulation **5085/2009** (2009) 93–101
11. Hamada, Y., Shimada, N., Shirai, Y.: Hand shape estimation using sequence of multi-ocular images based on transition network. In: Proceedings of the International Conference on Vision Interface. (2002)
12. Ramamoorthy, A., Vaswani, N., Chaudhury, S., Banerjee, S.: Recognition of dynamic hand gestures. Pattern Recognition **36** (2003) 2069–2081
13. Ahmad, T., Taylor, C., Lanitis, A., Cootes, T.: Tracking and recognising hand gestures, using statistical shape models. Image and Vision Computing **15** (1997) 345–352
14. Holden, E., Lee, G., Owens, R.: Australian sign language recognition. Machine Vision and Applications **16** (2005) 312–320
15. Tanibata, N., Shimada, N., Shirai, Y.: Extraction of hand features for recognition of sign language words. In: International Conference on Vision Interface. (2002) 391–398
16. Smith, P., da Vitoria Lobo, N., Shah, M.: Resolving hand over face occlusion. Image and Vision Computing **25** (2007) 1432 – 1448
17. Matthes, S., Hanke, T., Storz, J., Efthimiou, E., Dimiou, N., Karioris, P., Braffort, A., Choisier, A., Pelhate, J., Safar, E.: Elicitation tasks and materials designed for dicta-sign's multi-lingual corpus. International workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC), Valletta, Malte (2010) 158–163
18. Vassili, V.V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: in Proc. Graphicon-2003. (2003) 85–92
19. Kovac, J., Peer, P., Solina, F.: Human skin color clustering for face detection. In: EUROCON International Conference on Computer as a Tool. Volume 2. (2003) 144–148