Apart: Promoting the Next Generation of Digital Scholarship http://www.clir.org/pubs/reports/pub145/pub145.pdf

Galina, I. (2011). *El papel de las bibliotecas en las Humanidades Digitales*, 77th IFLA General Conference and Assembly, San Juan, Puerto Rico. 13-18 August 2011 http://conference.ifla.org/past/ifla77/104-russell-en.pdf

Galina, I., and E. Priani (2011*). Is There Anybody Out There? Discovering New DH Practitioners in other Countries. Digital Humanities 2011, Conference abstracts*. Stanford 2011, pp. 135-138

Revista Digital Universitaria-RDU (2011). *Las Humanidades Digitales*, Special Issue 12(7) http://www.revista.unam.mx/vol.12/num7/art68/index.html

Rheingold, H. (1993). *http://www.rheingold.com/vc/book/intro.html The Virtual Community*. Reading, Mass.: Addison-Wesley.

Presner, T., and C. Johanson (2009). *The Promise of Digital Humanities*. UCLA.

Notes

1. Known as *diplomado* this type of academic course is highly specialized and most cover a minimum of 120 hours.

Adaptive Automatic Gesture Stroke Detection

Gebre, Binyam Gebrekidan
binyamgebrekidan.gebre@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Wittenburg, Peter
peter.wittenburg@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

1. Introduction

Many gesture and sign language researchers manually annotate video recordings to systematically categorize, analyze and explain their observations. The number and kinds of annotations are so diverse and unpredictable that any attempt at developing non-adaptive automatic annotation systems is usually less effective. The trend in the literature has been to develop models that work for average users and for average scenarios. This approach has three main disadvantages. First, it is impossible to know beforehand all the patterns that could be of interest to all researchers. Second, it is practically impossible to find enough training examples for all patterns. Third, it is currently impossible to learn a model that is robustly applicable across all video quality-recording variations.

To overcome the three problems and provide practically useful solutions, this paper proposes a case-by-case user-controlled annotation model. The main philosophy for this kind of model is that a model designed to give the best average performance in a variety of scenarios is usually less accurate and less adaptable for a particular problem than a model tailored to the characteristics of that problem. This approach is also grounded in the 'No Free Lunch' theorems, which establish that for any algorithm, any elevated performance over one class of problems is offset by performance over another class (Wolpert & Macready 1997).

We apply our proposed solution to the problem of gesture stroke detection. To be more precise, for gesture stroke detection to be more accurate and more robust, we develop a model that takes intuitive input from the user for a given video and then we apply standard algorithms optimized to the characteristics of the video.

2. Gesture stroke

Gesture stroke is the most important message-carrying phase of the series of body movements that constitute a gesture (or the phrases in a gesture). The body movements usually include hand and face movements. The relevant questions for automatic stroke recognition are: a) what is a gesture? b) where does a gesture start and end? c) what are the phases in the gesture? d) which one is the stroke?

The literature does not give completely consistent answers to the above questions (Kendon 1980, 1972; Kita et al. 1998). However, the most prominent trend is that a gesture unit consists of one or more gesture phrases, each consisting of optional preparation, optional pre-stroke hold, obligatory stroke, optional post-stroke hold and optional retraction (Kendon 1980). Figure 1 shows the different phases in a gesture unit as outlined by Kendon.

Figure 1: Gesture phases (Kendon 1980, 1972)

For the purpose of this paper, a gesture phrase is classified into two: strokes and non-strokes. The non-stroke gesture phases include the absence of movement, the preparation, the hold, the retraction and any other body movements excluding the strokes.

3. Methodology

Our approach to determining gesture strokes involves four stages: a) detect face and hands for every person b) track them c) extract features d) distinguish strokes from non-strokes.

Different algorithms are used to solve each stage. Two features, corners and skin colors, are used to detect faces and hands. These features have been selected because they are usually stable from frame to frame for a given video.

Corners are shown to be good features for tracking (Shi & Tomasi 1993). They have the property that they are different from their surrounding points. A given point in a homogenous image cannot be identified whether or not it has moved in the subsequent frame. Similarly, a given point along an edge cannot be identified whether or not it has moved along that edge. However, the movement of a corner can conveniently be computed and identified, as it is non-ambiguous in its identity. This makes it a good feature for tracking.

For a given application, not all corners in a video frame are equally important. For gesture analysis, the interesting corners are the ones resulting from the body parts, mainly from face and hands. In order to filter out the corners irrelevant to body parts, we mask out corners that do not correspond to the skin color model.

The skin color model is developed with the involvement of the user. The user selects a representative instance of the skin color in one of the frames of the video, usually the first frame. And then the system extracts color information from that instance and finds all points in the frame where the matching of colors with the extracted color is high.

It is important to notice that the on-line selection of the skin region avoids having to design a skin color model for all human races. Off-line skin color model design is as practically difficult as collecting pictures of all human skin colors and its use in the detection for a particular skin color in a given video will be less accurate.

Given the corners from regions of the skin in the video, the tracking is done with the pyramidal implementation of the Lucas-Kanade algorithms (Bouguet 1999; Bradski & Kaehler 2008). Values extracted from the number of corners, clusters and their dynamics across frames (context) are fed into a supervised learning algorithm with class labels 1 for frames inside a stroke and 0 for frames outside a stroke. For the learning algorithm, we used support vector machine with RBF kernel.

4. Experiment data

Various videos have been used to test the detection of face, hands and their movements. Particularly, we experimented with two videos, each of which consists of two people of the same color: black and white skin colors. The resolution of the video with white people is 320x240 and that of the other video with black people is 1280x720. The higher the resolution, the better the detection quality.

For the supervised learning algorithm, we used a stroke and non-stroke annotated video data of 36 seconds long. This is the same video referred to above that has two white people speaking and gesturing. It has 914 frames, 847 of which are annotated. It has 19 strokes each ranging from 4 to 35 frames with mean 13.5 and standard deviation 6.5. It also has 40 non-stroke regions, which include moments of silence, preparation, retraction and holds.

All the videos for the reported experiments in this paper have been taken from the MPI archive `http://corpus1.mpi.nl/ds/imdi_browser/`.

A screen shot of a video with moving corners shown in blue. The rectangle is the region selected by the user as a model for the skin color (white)

Figure 3: A screen shot of a video showing detected skin regions. Very white regions correspond to high probabilities for white skin regions

Figure 4: A screen shot of a video with moving corners shown in blue. The rectangle is the region selected by the user as a model for the skin color (black)

Figure 5: A scrren shot of a video showing detected skin regions. Very white regions correspond to high probabilities for black skin regions

5. Results

The accuracy of results for face and hands detection based on initializations of regions of skin color on one frame in the video and applied to the subsequent frames in the same video shows that this approach can be very effective in identifying the corners belonging to the moving hands/face. Figures 2 through 5 show screen shots of the process and the results for a chosen video frame.

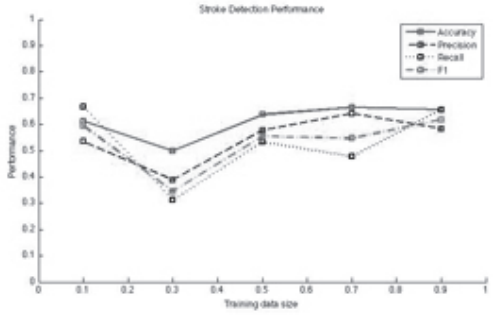

Figure 6: Stroke detection performance varaiation as training data size increases

The stroke/non-stroke classification results are shown in figure 6. The x-axis shows the training data size, which is 75% of the dataset. The y-axis shows the performance of the classification on test data, which is the remaining 25% of the dataset. As we vary the training data size, we get on average a slightly improving performance. The average accuracy, precision and recall achieved are 61.51%, 54.67% and 52.89%. The F1 measure is 53.26%. The average baseline (i.e. classifying every frame as non-stroke) achieves 57.55% accuracy and 0% recall. The performance measures show that there is a lot of room for improvement.

Evaluation for accuracy of frame boundaries for strokes and non-strokes should not be as clear-cut as we assumed in our experiments. One or two frame misses or shifts are tolerable given that even humans do not accurately mark the correct boundary, if any. However, we did not consider that observation in our evaluation results.

6. Conclusion

In this paper, we have put more emphasis on a more adaptive case-by-case annotation model based on the idea that with a little more input from users and facilitated by more user-friendly interfaces, annotation models can be more adaptive, more accurate and more robust (i.e. effectively deal with digital diversity). We have tested our approach on problems of hands/face tracking and automatic stroke detection.

We have noticed that building a skin color model online for all human skin colors will not only make the model more complex but also less accurate when applied on any particular video. However, a model built online for a given video initialized by input from the user achieves higher performance at no more cost than the initialization.

The correct detection of the skin color in combination with feature extraction algorithms can be used to study human gestures. We have shown that unique features (i.e. corners) and their dynamics across frames can be indicative of the presence of strokes, the most meaningful phase of a gesture. In our future research, we will continue to improve the stroke detection performance using more features and learning algorithms. We will also apply our methodology to classify strokes according to their meanings. Success of this methodology will have important impact in human activity recognition from video files.

Funding

The research leading to these results has received funding from the European Commissions 7th Framework Program under grant agreement no 238405 (CLARA). Special thanks go to Saskia van Putten and Rebecca Defina for providing one of the videos used in the experiments reported in this paper.

References

Adelson, E. H., C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden (1984). Pyramid methods in image processing. *RCA engineer* 29(6): 33-41.

Bouguet J. Y. (1999). *Pyramidal implementation of the lucas-kanade feature tracker: description of the algorithm*. Intel Corporation, Microprocessor Research Labs, OpenCV Documents, 3.

Bradski, G., and A. Kaehler (2008). *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media.

Kendon, A. (1972). Some relationships between body motion and speech. *Studies in dyadic communication* 7:177.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication* 25: 207-227.

Kita, S., I. van Gijn, and H. van der Hulst (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. *Gesture and sign language in human-computer interaction*, pp. 23-35.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. U of Chicago P.

Shi, J., and C. Tomasi (1993). *Good features to track. IEEE computer society conference on computer vision and pattern recognition*, pp. 593-600.

Wolpert, D. H., and W. G. Macready (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation* 1(1): 67-82.