Human Behavior Analysis from Video Data Using Bag-of-Gestures

Víctor Ponce[1 2], **Mario Gorga**[1 2], **Xavier Baró**[1 2 3], **Sergio Escalera**[1 2]
[1]Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona.
Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain Barcelona, Spain.
[2]Centre de Visió per Computador, Campus UAB, Edificio O, 08193, Bellaterra, Barcelona.
[3]Universitat Oberta de Catalunya Rambla del Poblenou 156, 08018, Barcelona, Spain
v88ponce@gmail.com, {mgorga,xevi,sergio}@maia.ub.es

Abstract

Human Behavior Analysis in Uncontrolled Environments can be categorized in two main challenges: 1) Feature extraction and 2) Behavior analysis from a set of corporal language vocabulary. In this work, we present our achievements characterizing some simple behaviors from visual data on different real applications and discuss our plan for future work: low level vocabulary definition from bag-of-gesture units and high level modelling and inference of human behaviors.

1 Motivation

Feature extraction and gesture recognition from non-verbal language are of particular interest in the analysis of psychological factors that a subject presents [Winsor *et al.*, 1997]. In this work, we are interested in recognizing a large set of human gestures from video data.

The gesture recognition issue has been treated in recent years by several authors. Most approaches are based on Dynamic Time Warping (DTW) to perform a dynamic alignment in time of gesture features, and it has been successfully applied to Sign Language and speech recognition [Athitsos, 2010]. An extended version of DTW has been proposed in [Stefan *et al.*, 2008], where different feature candidates are tracked and dynamically aligned in order to reduce the influence of noise in images. Another general framework used to model temporal series and human behaviors is Hidden Markov Model (HMM) [Fang *et al.*, 2007]. In [Alon *et al.*, 2009] and [Fang *et al.*, 2007] two frameworks for gesture recognition are proposed, which take into account spatio-temporal matching and training of gesture models by means of DTW and HMM, respectively.

Using standard computer vision feature extraction and machine learning approaches, our work is based on the definition of a large vocabulary of gesture units, which we call Bag-of-Gestures (BOG), and a probabilistic modelling of human gestures. The next sections review our achievements and discuss our plan for future work.

2 Achievements

For the feature extraction stage of a gesture recognition system, we defined a set of simple visual features. These features are based on face detection, skin modelling, and feature tracking processes. We used the Face Detector of Viola & Jones [Viola and Jones, 2004] in order to detect the region of the face and define our origin or coordinates, which is similar to the approach presented in [Stefan *et al.*, 2008]. Inner pixels of the detected region are used to train a skill color model [Jones and Rehg, 2002], which is used to look for hand/arm candidate regions. Finally, those blobs connected with the highest density correspond to our regions of interest, which are tracked using mean shift [Fukunaga and Hostetler, 1975]. All the spatial coordinates of the detected regions are computed in reference to the face coordinates and normalized using the face area. Examples of detected and tracked regions are shown in Figure 1.

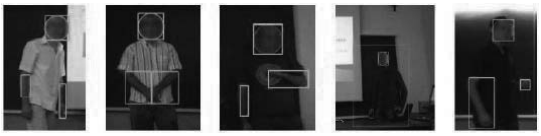

Figure 1: Examples of detected regions for different student presentations.

With the computed feature space, we performed an initial experiment where 15 bachelor thesis videos were recorded (some examples are shown in Figure 1). Using the social signal indicators defined in [Pentland, 2005], we computed a set of activity, stress, and engagement indicators from the extracted feature space [Ponce *et al.*, 2010]. Using the score obtained by the teachers at the presentations, we categorized the videos in two levels: those with the lowest score, and those with the highest score, and trained a Discrete Adaboost binary classifier [Friedman *et al.*, 1998]. Applying stratified ten-fold cross-validation, we obtained interesting results, showing high prediction performance of student score based on his/her non-verbal communication using the extracted features. Moreover, we used Adaboost margin in order to rank features by relevance [Ponce *et al.*, 2010]. Furthermore, we have also tested our system on different applications, such as Sign Language recognition using a novel multi-target dynamic gesture alignment, Attention Deficit Hyperactivity Disorder (ADHD), corporal physiotherapy analysis, and in-patient monitoring, with high success [Ponce *et al.*, 2010]. See some examples in Figure 2.

In the next section we discuss our plan to extend the presented methodology to define a set of gesture unit vocabulary

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

Figure 2: (a) Sign language recognition, (b) Inpatient monitoring, (c) Physiotherapy analysis, and (d) Attention deficit hyperactivity disorder analysis.

as well as to define a higher level recognition of human behaviors.

3 Future work

Future work involves the recognition of a large number of possible gestures, each one defined as a composition of gesture-units from a behavior vocabulary. To achieve these goals, we will perform the following steps: a) compute the feature-space from a large set of videos involving several gestures; b) compute bag-of-gesture units vocabulary (BOG); c) compute the training data in BOG representation and train a temporal model to estimate state transitions; d) final recognition will consist of feature extraction, BOG representation, and inference from trained temporal model on the new gesture vocabulary. Next, each step of the proposed framework will be described in detail.

3.1 Feature-space computation

Initially we plan to extract the features described previously and focus on the training and inference of gestures.

3.2 Bag-of-Gesture units

Gestures are usually described as a sequence of features per sample. Few works have decomposed gestures into a set of units [Alon *et al.*, 2009]. Our goal is to perform temporal clustering of gestures by means of DTW. Using this methodology we can categorize gesture units of different length based on a clustering cost. In the text recognition and visual categorization fields, Bag-of-Words and Bag-of-Visual-Words have been applied with successful results. Our plan is to extend the same idea to the definition of gesture units, which are deformed and aligned in time.

3.3 Temporal Model Training

The sequence of gesture units on the time can be modeled using a temporal model, such as HMM, often used in the literature. From the definition of a gesture vocabulary, a discrete alphabet λ is obtained. If we consider each symbol of the alphabet (a gesture unit) as a possible state of a gesture, we can train state-transition probabilities of a HMM with standard Baum-Welch algorithm. We also plan to train and propose different graphical model structures in order to improve gesture recognition performance. Although HMM are the classical used temporal models, and will be used to verify the framework, we plan to introduce more powerful models, such as the Dynamic Bayesian Networks (DBNs) [Rabiner, 1989], and a generalization of HMM and Linear Dynamical Systems (LDSs) by representing the hidden (and observed) state in terms of state variables, which can have complex interdependencies.

3.4 Inference

The final inference consists of the testing of the proposed methodology on a large scale data set of gestures. The process will be performed by quantifying a BOG vocabulary and doing inference on trained temporal models of gestures.

4 Acknowledgements

This work has been supported in part by the projects TIN2009-14404-C02, CONSOLIDER-INGENIO CSD 2007-00018 and 2009PID-UB/04.

5 Conclusion

We reviewed our achievements in the scope of human behavior analysis from video data: a) description of gesture features, b) new data set of videos, and c) real applications: Social signal analysis, Sign Language recognition using dynamic gesture alignment, ADHD, corporal physiotherapy analysis, and inpatient monitoring. Furthermore, we discussed our future lines of research: a) definition of a Bag-of-Gesture vocabulary. BOGs are computed using DTW, obtaining an alphabet of gesture primitives of different length, b) a higher level behavior analysis training graphical model, and c) its use on large scale and real challenging health-care applications.

References

[Alon *et al.*, 2009] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *PAMI*, 31(9):1685–1699, 2009.

[Athitsos, 2010] V. Athitsos. Large lexicon project: Asl video corpus and sl indexing/retrieval algorithms. *RPSL: Exploitation of Sign Language Corpora*, 2010.

[Fang *et al.*, 2007] G. Fang, W. Gao, and D. Zhao. Large-vocabulary continuous sign language recognition based on transition-movement models. *SMC-A*, 37(1), 2007.

[Friedman *et al.*, 1998] J. Friedman, T. Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:2000–2030, 1998.

[Fukunaga and Hostetler, 1975] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

[Jones and Rehg, 2002] M. Jones and J. Rehg. Statistical color models with application to skin detection. *IJCV*, 46:81–96, 2002.

[Pentland, 2005] A. Pentland. Socially aware computation and communication. *Computer*, 38:33–40, 2005.

[Ponce *et al.*, 2010] V. Ponce, S. Escalera, X. Baro, and P. Radeva. Automatic analysis of non-verbal communication. *CVCRD10 Achievements and New Opportunities in Computer Vision*, pages 105–108, 2010.

[Rabiner, 1989] L. R. Rabiner. A tutorial in hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

[Stefan *et al.*, 2008] A. Stefan, V. Athitsos, J. Alon, and S. Sclaroff. Translation and scale invariant gesture recognition in complex scenes. *PETRA*, 2008.

[Viola and Jones, 2004] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[Winsor *et al.*, 1997] J. L. Winsor, D.B. Curtis, and R.D. Stephens. National preferences in business and communication education. *JACA*, 3:170–179, 1997.