Research Statement

Karan Sikka
Machine Perception Lab
University of California San Diego

January 22, 2013

The field of Computer vision has witnessed significant improvement over the last two decades both with progress in machine learning and improvement in computing power. Hence it has been possible to employ computer vision to real-life applications such as face detection and recognition, and object matching. An interesting venue that has emerged since 90's is application of machine vision to model and infer human facial expressions. Facial expressions are important since they provide a wealth of information about a person's internal states like emotion, intention, etc. This high-level information could be useful for numerous tasks such as patient monitoring in clinical application, automatic student tutoring, and social robotics. Early attempts at facial expression recognition were targeted towards **posed** expressions, mostly in highly controlled environment. However it took sometime for researcher to design vision algorithms that could identify **spontaneous** expressions in more naturalistic conditions. Current work is particularly focused towards handling the challenges associated with spontaneous facial expression recognition.

Spontaneous expressions differ from posed expressions since they are produced by different parts of the brain [1]. Thus they contrast in both which muscle are moved and dynamics of the movement. From the viewpoint of computer vision spontaneous expressions involve challenges like out-of-plane head movements (non-frontal face), self-occlusion, accompanied verbal communication, among others. Previous studies [2] have highlighted that an efficient model for inferring spontaneous facial expressions should incorporate temporal dynamics in a video signal effectively. Encoding temporal dynamics in videos is however not an easy task and is plagued by two major challenges: (1) video signals of varied length, (2) lack of apriori knowledge about the time point and duration of the expression of interest. One possible solution is manual labeling at frame level (compared to sequence level labels), however this could be expensive, labor intensive and prone to errors. Thus identifying facial expressions in videos with limited ground-truth poses interesting challenges and opens up new research venues.

Our starting point consists of exploring an approach where we model the problem of expression classification in videos with limited ground-truth as a weakly supervised classification problem. Weakly supervised classification is

relatively new paradigm designed for cases where the training dataset is incomplete or ambiguous with respect to the problem, referred to as weakly-labeled data. For instance, a training dataset is weakly labeled in an object detection problem if it merely informs about the presence/absence of an object without any location information. Yet we are also interested in the possible location of the object during inference. In our recent work on pain classification in videos [3] (appearing in FG'13) the ground-truth data only informed about presence/absence of pain in entire sequence. It employed a latent model (called multiple instance learning) for learning and prediction combined with a novel way of representing each video as bags of multiple segments (or sub-sequences). The bags of segments were built using discriminative Bag of Words architecture proposed by us for facial expression classification in ECCV'12 workshop [4] called 'What's in the Face'. The multiple segment representation allowed us to handle non-frontal poses common in spontaneous expressions and also incorporate temporal information in each segment. Our pain classification algorithm was able to address a number of limitations in earlier approaches: (1) handling ambiguity introduced by sequence level ground-truth, (2) incorporating temporal dynamics when there is uncertainty about the duration, extent and number of occurrences of pain signal, and (3) providing a more intuitive learning and prediction model. This model was able to achieve state-of-the-art results on a publically available pain dataset. However the performance numbers were still far from perfect and left many unanswered research questions.

Currently we are investigating the application of various latent models for the problem of spontaneous facial expression recognition in video. This involves choosing apt underlying assumptions for a particular problem and also making the inference and learning procedures tractable. We also believe that the analysis of the underlying patterns discovered by models involving latent variables will be of significant interest to the community.

References

[1] A. Miehlke, U. Fisch, and C. Eneroth, *Surgery of the facial nerve.* Urban & Schwarzenberg Munich, 1973.

[2] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[3] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain classification using multiple instance learning," in *Automatic Face & Gesture Recognition and Workshops (FG 2013), 2013 IEEE International Conference on.* IEEE, 2013.

[4] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring bag of words architectures in the facial expression domain," *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 250–259, 2012.