CROWD ANALYTICS VIA ONE SHOT LEARNING AND AGENT BASED INFERENCE

Peter Tu, Ming-Ching Chang, Tao Gao

GE Global Research

ABSTRACT

For the purposes of inferring social behavior in crowded conditions, three concepts have been explored: 1) the GE Sherlock system which makes use of computer vision algorithms for the purposes of the opportunistic capture of various of social cues, 2) a one shot learning paradigm where behaviors can be identified based on as few as a single example and 3) an agent based approach to inference where generative models become the basis for social behavior recognition. The Sherlock system makes use of tracking, facial analysis, gaze estimation and upper body motion analysis. The one-shot learning paradigm makes use of semantically meaningful affects as descriptors. The agent based inference methods allows for the incorporation of cognitive models as a basis for inference.

Index Terms— Tracking, Expression, Agent, Inference, Learning, Crowds

1. INTRODUCTION

This paper describes a variety of methods that have been developed for the purposes of understanding crowd level behaviors using stand-off video analytics methods. Three main topics are considered: 1) the GE Sherlock System: a comprehensive approach to capturing and analyzing non-verbal cues of persons in crowd/group level interactions, 2) One Shot Learning: a new approach to crowd level behavior recognition based on the concept that a new behavior can be recognized with as little as a single example and 3) Agent Based Inference: a novel approach to the analysis of individual cognitive states of persons interacting in a group or crowd level context. The paper starts with a description of the GE Sherlock system which encompasses methods such as person tracking in crowds, dynamic PTZ camera control, facial analytics from a distance such as gaze estimation and expression recognition, upper body affective pose analysis and the inference of social states such as rapport and hostility. The paper then discusses how cues derived from the Sherlock system can be used to construct semantically meaningful behavior descriptors or affects allowing for signature matching between behaviors which can be viewed as a form of one shot learning. Going beyond affects based on direct observation, we argue that more meaningful affects can be constructed via the inference of the cognitive states of each individual. To this end we introduce the Agent Based Inference framework.

This work is jointly supported by (1) grant award 2013-IJ-CX-K010, National Institute of Justice, US Department of Justice and (2) the Defense Advanced Research Projects Agency (Contract D13PC00002). The information presented here does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

2. SHERLOCK

The GE Sherlock system is based on the hypothesis that by jointly considering a wide variety of visual cues such as facial expression, gaze direction, body posture and motion, it is possible to estimate complex group level social states in a completely automated fashion. It is argued that social interaction analysis can be cast as a latent variable inference problem. The system operates over multiple individuals functioning freely in unconstrained environments with no person borne instrumentation.

A real time stand-off end-to-end social interaction analysis system has been instantiated. Individuals can move freely while being tracked by a set of fixed RGB+D cameras, which produce estimates of location and articulated body motion. A ring of PTZ cameras are tasked based on such tracking results to capture high resolution facial imagery. Facial landmark fitting and tracking is performed so as to extract facial expressions and gaze directions. The real-time social interaction system distills the stream of person specific cues into a set of site-level aggregate statistics which are independent of the configuration and number of observed individuals. Such measures include: emotional affect, (derived from observed facial expressions), proximity (derived from tracked positions), activity/motion (derived from body motions) and engagement (derived from position and gaze direction). The system continuously produces these statistics resulting in a time-series representation. Sets of graphical models operate over these measures resulting in a continuous estimate of various group-level social states such as rapport and hostility.

In terms of system architecture, a modular design was instantiated where by each component consumes necessary inputs such as raw video feeds and meta-data generated by other modules and in turn each module produces meta-data that is inserted into a message-passing publish and subscribe architecture. Using multiple computing platforms, a real time system which includes: multi-camera tracking, PTZ control, facial analysis, data-consolidation and social-state inference

has been constructed. This type of modular design allows for the incorporation of multiple third party capabilities into this system of systems computer vision architecture. The following are a set of brief technical descriptions of the core video analytic modules.

Tracking: A detect-and-track paradigm is used to estimate the location and trajectory of each subject that are located in a specified region of interest. Multiple RGB(D) cameras are initially calibrated with respect to a world coordinate system. Imagery from each camera is used to independently produce a set of person detections and associated appearance signatures. These detections are then matched to existing trackers. Detections that are not associated with an existing tracker can be used to initialize a new tracker. Trackers that persistently fail to be associated with new detections are terminated.

Articulated Motion Analysis: In addition to tracking, the RGB(D) camera imagery are used to extract motion cues known as "space-time-corners". These cues are associated with a spacial histogram defined based on the measured location and height of each subject. These spatial/frequency distributions are used as a representation of articulated motion body based on RGB imagery captured with the PTZ cameras.

PTZ Camera Control: The location of each PTZ camera is initially measured with respect to the world coordinate system. A calibration procedure is used to map pan (P), tilt (T) and zoom (Z) values to (X,Y,Z) coordinates in the world coordinate system such that if a face is located at (X,Y,Z) then the resulting imagery from the PTZ camera will allow for various forms of facial analysis. The tracking system produces the location of each person in ground plane coordinates (X,Y). The Z value is determined based on an estimate of subject height. An optimization algorithm is used to automatically assign PTZ cameras to tracked subjects.

Facial Analysis: Given high resolution imagery produced by the PTZ cameras, the following steps are taken: i) face detectors are used to produce a bounding box of the subject's face, ii) eye detectors are used to locate the subject's eyes, iii) if both eyes are detected, a facial landmark model is fitted to the subject's face, iv) an estimate of the vertical and horizontal gaze directions are computed based on the shape of the fitted landmark model, v) an estimate of the horizontal eyeball location is computed allowing for detection of events such as an "averted gaze", vi) the fitted landmark model is used to synthesize a frontal view of the subject's face, vii) gross facial expression models are used to estimate a set of common facial expressions.

Inference: Given a stream of meta-data associated with each subject (location, articulated motion, gaze direction, facial expression) a set of aggregate social signals are produced. For the purposes of inferring group level social concepts such as rapport and hostility, graphical models are used to reason over the aggregate social signals resulting in real-time estimates of the probability distribution associated with each social concept.

3. ONE SHOT LEARNING

Given a single observation of an instance of a query behavior (such as group formation), the recognition system must be able to classify any subsequent observations as either being another example of this class or not. From a recognition perspective this is similar to the problem of face recognition where at any time the system is given two faces, the system must then decide whether or not the two face images originated from the same or different individual(s). Similarly a one shot behavior recognition system must take any pair of observed behaviors and determine whether or not the two behaviors are the same or not.

Approach: we start by defining the following terms:

- Behaviors: are sequences of events that are performed by people.
- Video Analytic Streams: data that is generated by base video analytics such as the location, gaze direction, expression and motion field of each person (we use the previously described GE Sherlock system).
- Signal Generators: various interpretations of the data such as the observation of events as well as measured quantities (these are hand crafted). Can be viewed as a single variable time series between 0 and 1.
- Affects: these are semantically meaningful descriptors of behaviors; a signal generator must be able to produce a single affect score between 0 and 1.
- Signatures: these are structures used to characterize a behavior. They encode the observed affects. We also consider the time at which each affect was observed with the idea that the sequence of affects may be important.

A signal generator is a module that a) will consume a video analytic stream, b) must analyze this stream so as to produce a time series with values ranging from 0 to 1, c) at the completion of each behavior must be able to produce an affect score between 0 and 1 and d) must have a set of parameters that define the behavior of this module. By allowing for a parametric representation for each signal generator, a user can instantiate a particular variant of a given signal generator. Conversely, multiple variants of a signal generator can be produced by considering various permutations of the signal generator parameters. A signal generator bank maintains the set of signal generators that will be used to characterize a given behavior.

Once a behavior has been processed by a signal generator bank, the signature is constructed that allows for the description of a given behavior. At this time, such a signature is simply the affect scores generated by each signature generator.

Matching between a pair of signatures is achieved by considering a weighting of affect differences. The determination of affect weights is achieved via an optimization process.

Experiments involved the collection of 13 behavior pairs and the instantiation of an initial set of 18 signal generators. Pairwise match scores were calculated using the following distance scores:

$$d(b_k, b_j) = \sum_{i=0}^{n} w_i ||sg_i(b_k) - sg_i(b_j)|| \quad (1)$$

Where d is a distance measure, b is an observed behavior, sg is the affect score for a given signal generator, n is the number of signal generators and w is a weight associated with each signal generator. Matching results were computed for both a uniform weighting function as well as an optimized weighting function - see Figure 1.

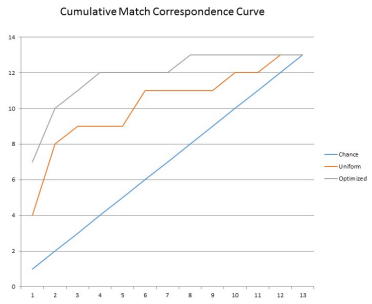

Fig. 1. A Cumulative Match Characteristic (CMC) curve for the one shot learning experiments. The y axis defines the number of true matches that receive a rank of x or better. Blue curve: performance that would be expected using chance alone. Red Curve: performance observed using a uniform weighting function (all signal generators contributing equally). Green: performance observed using an optimized weighting function.

4. AGENT BASED INFERENCE

Video analytics allows for the direct observation of cues such as position, affective pose, gaze direction, expression and gesture. These measurements can be used to characterize the physical states of individuals participating in group level behaviors. However, such individuals must be seen as cognitive agents that are in possession of a subjective form of internal experience. Models for these internal experiences may include concepts such as: emotions, intentions, goals, plans, expectations and representations of other individuals (theory of mind).

With this concept in mind each individual can be viewed as an agent where: 1) each agent has a physical state that is observable and a latent internal state that is not open to direct observation, 2) the internal state drives the physical state, 3) future internal states are based on the current internal state as well as observations of the physical state of individuals interacting with the agent. These relationships are depicted in Figure 2.

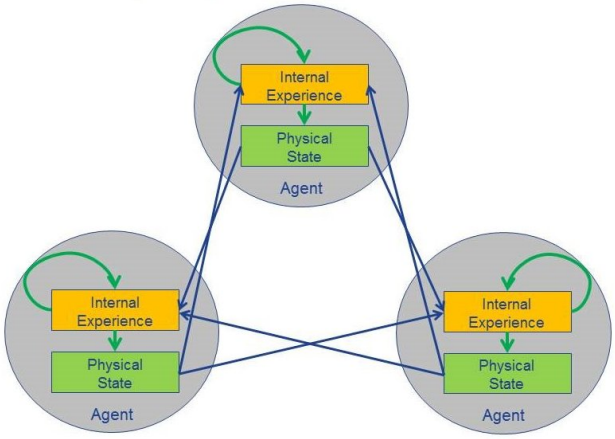

Fig. 2. This figure depicts three interacting agents, each with their own internal and physical states. The physical state of each agent is driven by the internal state. The internal states evolve as a function of their current internal states as well as the observations of the physical states of other interacting agents.

As previously described, the affects used for characterizing group level behaviors have to date been based on direct observations of the physical states of each individual. If, based on such observations, estimates of the internals states of each agent can be made; we argue that a much richer set of affects can be derived. However, due to the non-linear and stochastic nature of these processes, it will not be possible to compute such latent variables via direct inversion of the observations. We now consider inference strategies based on the ability to model and hence simulate agent based behavior.

Given a sequence of observed physical states of a group of interacting individuals, the task at hand is to infer the corresponding sequence of latent internal states. If the latent and observable states can be characterized by a set of symbols, then an equivalent task is the computation of the full sequence of interlaced latent and observed symbols given the observed symbol sequence. By modeling the mechanisms associated with the internal states, agent based simulators can be used to synthesize complete latent/observable behavior sequences. A potentially infinite set of such synthetic sequences can then be produced via random sampling methods. It can be argued that agent based simulation allows for two types of inference strategies: ”Hypothesis and Test” and ”Recognition via Machine Learning”.

The hypothesize and test approach is based on the idea

of 1) synthesizing a large number of possible behavior sequences, 2) developing a similarity measure that can compare any two sequences based on physical symbols alone, and 3) estimating the latent symbols of the query sequence based on the latent symbols of the most similar synthetic latent sequence. As previously stated the cardinality of the possible synthetic sequence set is potentially infinite and so nave approaches such as approximate nearest neighbors are not tractable. We thus consider the use of Multiple Hypotheses Tracking (MHT) as an appropriate method for this form of inference. In particular we propose the use of a particle filtering framework.

Particle filtering is an iterative process which attempts to estimate the temporal evolution of a set of latent state variables given an observation sequence. At time 0 an initial set of particles are randomly instantiated. Each particle consists of an estimate of the initial latent variable values and a prediction of the associated observable variables. The likelihood of each particle can be computed based on the similarity of the predicted and observed physical states. Sampling methods are then used to nominate particles for propagation to the next iteration based on these likelihood measures. Particle propagation is based on stochastic sampling. In this way particles that are able to track the observation sequence are allowed to persist. Particles that fail to predict the observed behavior are subsequently culled. The output of this process is thus the most likely interpretations of the query sequence.

Like all forms of Multiple Hypothesis Tracking, Particle Filtering is a form of search that is reliant on accurate modeling of system dynamics. As the complexity of the internal experience models increases, so too does the fear of being trapped by local minima in the search space. We thus argue that instead of propagation via random sampling, we consider the use of recognition methods as a mechanism for guiding the evolution of the particle filters. To this end we propose the use of Recurrent Neural Networks (RNNs).

Originally developed for natural language processing (NLP), RNNs can be viewed as a form of symbolic sequence recognition. For the purposes of illustration, consider the entire works of William Shakespeare. Each word can be represented by a unique symbol. Each sentence can then be viewed as a symbolic sequence. The corpus of Shakespearean plays thus becomes training data. Once an RNN has been trained, it can be given an initial seed sequence with as few as a single symbol. The RNN will then produce a probability distribution for the next element in the sequence. Sampling methods can then be used to select the next element. This process can be repeated multiple times resulting in the synthesis of complete sequence that will appear to resemble the properties of the training data. For example, given an initial seed of "The dog", a Shakespearian RNN might produce the following sentence: "The Dog cometh from yonder castle". We propose the following:

1. Agent based simulators be used to construct a corpus of training data needed for construction of a behavior RNN.

2. Stochastic sampling on appropriately trained RNNs will be incorporated into the particle filtering framework. Instead of each particle having its own generative internal experience models, the particle will have a RNN.

3. The particle RNNs will initially be seeded with random internal symbols. They will then predict through sampling the next set of physical symbols. These predicted physical symbols will be compared with the physical symbols of the query sequence. Likely particles will be allowed to transition to the next iteration. In this case transitioning involves predicting the next set of internal symbols.

We argue that the merit of this approach revolves around the idea that while there are a potentially infinite number of possible behavior sequences, by and large the majority of sequences that will be encountered can be associated with a relatively small number of behavior modes. Given appropriate training data knowledge associated with such modes will be encapsulated by the RNNs. Success of this inference paradigm is predicated on the ability to produce high fidelity cognitive agent based simulators.

In terms of implementation, cognitive models have been developed which incorporate concepts such as: character types, internal emotions and policy driven interactions. Methods for converting observed and generated behaviors into symbolic sequences have been constructed. An RNN has been trained based on symbol sequences derived from forward simulations. Using simulated data with associated ground truth information, it has been demonstrated that particle filtering is able to infer latent variables associated with the cognitive models.

5. RELATION TO PRIOR WORK

A large number of papers have focused on details regarding the extraction of socially relevant cues based on computer vision methods - see [1, 2] for a survey of such methods. To our knowledge, the Sherlock system stands out due to the opportunistic methods used for operation in crowded environments. Going beyond computer vision methods, there exists a wide body of work focused on developing models for the interpretation of social cues [3, 4, 5, 6, 7, 8, 9, 10]. Adding to this body of work, this paper focuses on the methods of inference as opposed to specific models. Prior work with respect to the Sherlock system itself can be found in [11, 12]. From an analysis perspective, this paper expands the concept of inference to include semantic affects and cognitive models.

6. REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," in *Image and Vision Computing*, 2009, vol. 27, pp. 567–578.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions.," in *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 2009, vol. 31(1), pp. 39–58.

[3] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, "Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2001, pp. 301–308.

[4] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze.," in *Pattern Analysis and Machine Intelligence, IEEE Transactions On*, 2010, vol. 32(3), pp. 487–500.

[5] D. Metaxas and S. Zhang, "A review of motion analysis methods for human nonverbal communication computing. image and vision computing,," in *Image and Vision Computing*, 2013, vol. 31(67), pp. 421–433.

[6] J. E. Grahe and F. J. Bernieri, "The importance of nonverbal cues in judging rapport," in *Journal of Nonverbal Behavior*, 1999, vol. 23(4), pp. 253–269.

[7] F. J. Bernieri, J. S. Gillis, J. M. Davis, and J. E Grahe, "Dyad rapport and the accuracy of its judgment across situations: A lens model analysis," in *Journal of Personality and Social Psychology*, 1996, vol. 71(1), p. 110.

[8] C. Goodwin, "Conversational organization: Interaction between speakers and hearers," in *New York: Academic Press*, 1981.

[9] L. Mol, E. Krahmer, A. Maes, and M. Swerts, "Adaptation in gesture: Converging hands or converging minds?," in *Journal of Memory and Language*, 2012, vol. 66(1), pp. 249–264.

[10] A. Kendon, "Gesture: Visible action as utterance," in *Cambridge University Press*, 2012.

[11] P. Tu, J. Chen, M. Chang, T. Yu, T. Tian, G. Rubin, J. Hockett, and A. Logan-Terry, "Cross-cultural training analysis via social science and computer vision methods," in *6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015)*, 2015.

[12] J. Chen, M. Chang, T. Tian, T. Yu, and P. Tu, "Bridging computer vision and social science : a multi-camera vision system for social interaction training analysis," in *International Conferenence on Image Processing*, 2015.