Social Interactive Human Video Synthesis

Dumebi Okwechime, Eng-Jon Ong, Andrew Gilbert, and Richard Bowden

CVSSP, University of Surrey, Guildford, Surrey, GU17XH, UK
{d.okwechime, e.ong, a.gilbert, r.bowden}@surrey.ac.uk

Abstract. In this paper, we propose a computational model for social interaction between three people in a conversation, and demonstrate results using human video motion synthesis. We utilised semi-supervised computer vision techniques to label social signals between the people, like *laughing*, *head nod* and *gaze direction*. Data mining is used to deduce frequently occurring patterns of social signals between a speaker and a listener in both *interested* and *not interested* social scenarios, and the mined *confidence* values are used as conditional probabilities to animate social responses. The human video motion synthesis is done using an appearance model to learn a multivariate probability distribution, combined with a transition matrix to derive the likelihood of motion given a pose configuration. Our system uses social labels to more accurately define motion transitions and build a texture motion graph. Traditional motion synthesis algorithms are best suited to large human movements like walking and running, where motion variations are large and prominent. Our method focuses on generating more subtle human movement like head nods. The user can then control who speaks and the interest level of the individual listeners resulting in social interactive conversational agents.

1 Introduction

Human motion synthesis has extensive applications in the movie and gaming industry. The ability to control the movement of a character in a video scene can provide an attractive alternative to post-production filming, providing movie editors with the means to edit an actor's performance without having to re-record the scene, which can be very expensive and time consuming. Although CGI (Computer Generated Imagery) is the vastly popular medium for computer game characters, photo-realistic human animation has proven to heighten realism in the gaming experience especially in combat platforms. Though there has been extensive research in the field for motion capture and human video texture synthesis, little work has been done in developing video based socially interactive avatars, capable of responding appropriately to non-verbal communication. This issue is addressed in this paper.

Our aim is to develop a human video *Motion Model*, specifically tailored for synthesising social interactive behaviour. This is done by combining a Probability Density Function (PDF) with a Markov Transition Matrix to derive the likelihood of pose and the probability of transitions respectfully. To increase reliability of the motion model, we introduce *Texture Motion Graph*, akin to Motion

Graphs by Kovar et al [1]. We extend the approach to allow multiple identical subgraphs to increase connectivity.

We propose a novel approach to derive social dynamics using data mining [2] to efficiently identify social trends between a group of three people in a conversation. The *confidence* values extracted from the mining are used as weighted conditional probabilities to generate appropriate non-verbal responses, resulting in a fully automated social interactive system. The user can select who speaks and subsequently who listens, and control the listeners level of interest in the conversation, effectively changing their social dynamics. Although the control of social dynamics is demonstrated on video synthesis, the elements of the computational model of interaction (i.e. *gaze*, *nod*, *laugh* etc) can be readily extended to other synthesis approaches such as *motion captured skeleton* animations.

The paper is divided into the following sections. Section 2 briefly details a background in the field of motion texture synthesis. Section 3 presents an overview of the entire system. Section 4 and 5 describes the approach of generating the motion model and deriving trends in the social dynamics respectfully. Section 6 presents the social interaction motion control, and the remainder of the paper describes the results and conclusion.

2 Background

Synthesis has extensive applications in graphics and computer vision, and can be categorised into three groups: textures synthesis of discrete images, temporal texture synthesis in videos, and motion synthesis in motion captured data. Early approaches to texture synthesis were based on parametric [3] and non-parametric [4] methods, which create novel textures from example inputs. Kwatra et al [5] generate perceptually similar patterns from a small training data set, using a graph cut technique based on Markov Random Fields (MRF). Approaches to static texture synthesis paved the way for temporal texture synthesis methods, often used in the movie and gaming industries for animating photo-realistic characters and editing video scenery. An example is presented by Bhat et al [6] who used texture particles to capture dynamics and texture variation travelling along user defined flow lines. This was used to edit dynamic textures in video scenery.

A number of researchers have used statistical models to learn generalised motion characteristics for synthesis of novel motion [7] [8] [9]. Unfortunately all these systems use a generalisation of the motion rather than the original data, and cannot guarantee that the synthesised motion looks natural.

Motion synthesis using example-based methods, i.e. retaining the original motion data to use in synthesis, provides an attractive alternative as there is no loss of detail from the original data [10] [11] [12] [13]. Representing motion transitions using a *motion graph* [14] [15] [16] [17], originally introduced by Kovar et al [1], provides additional user-control on positioning, using both pieces of original data and automatically generated transitions to perform an optimal graph walk that satisfies user-defined constraints. Treuille et al [13] developed a system that synthesizes kinematic controllers which blend subsequences of precaptured motion clips to achieve a desired animation in real-time.

In some cases, techniques used for motion synthesis of motion captured data, are similar to the techniques used for temporal texture synthesis of videos. By substituting pixel intensities (or other texture features) with marker coordinates, and applying motion constraints suited to the desired output, a similar framework can be extended to both domains. Schödl et al [18] introduced *Video Textures* which computes the distances between frames to derive appropriate transition points to generate a continuous stream of video images from a small amount of training video. Similarly, Flagg et al [19] presented *Human Video Textures*, where, given a video of a martial artist performing various actions, they produce a photo-realistic avatar which can be controlled, akin to a combat game character. In these cases, human texture synthesis is performed on periodic data, or constrained to guarantee the actor returns to a neutral pose.

Research in social interaction can be grouped into two main categories: emotion based on cognitive psychology [20], and linguistics based on dialogue understanding [21]. Though emotion understanding using high level deduction of social behaviour, like tone of voice and facial expressions, is of vital importance in how people socially interact, emotion recognition in a natural conversation, especially amongst adults, is a very complex problem and would require extensive data and research in deducing social trends. Also, structured dialogue can not be easily interpreted to observe generalised unconscious and non-verbal social behaviour.

3 Overview

Our proposed system consists of two stages: **Human Video Texture Synthesis**, and a **Social Interaction Model**.

Video Texture Synthesis allows a user to reproduce motion in a novel way by specifying which type of motion inherent in the original sequence to perform. Given a data set of a full body video sequence, following dimensional reduction via PCA, an unsupervised segmentation derives cut point clusters, where each cluster represents groups of similar frames that can seamlessly blend together. A texture motion graph is built to guarantee connectivity. Finally, a dynamic model is learnt, combining kernel density estimation with a markov transition matrix to derive the likelihood of transitioning from one cut point to another to generate novel motion sequences.

Developing a *Social Interaction Model* starts with a social behaviour experiment, whereby scenarios when the listener is *interested* and *not interested* in the topic of conversation can be determined. Using various action recognition techniques, social signals of the video are labelled such as *head nods*, *laughing* etc. Data mining is used to derive trends between the listener and the speakers given these social signals, producing conditional probabilities of a listener's social behaviour given a speaker's social signal. These conditional probabilities are used as weights to drive the motion model, given the user control over the social dynamics.

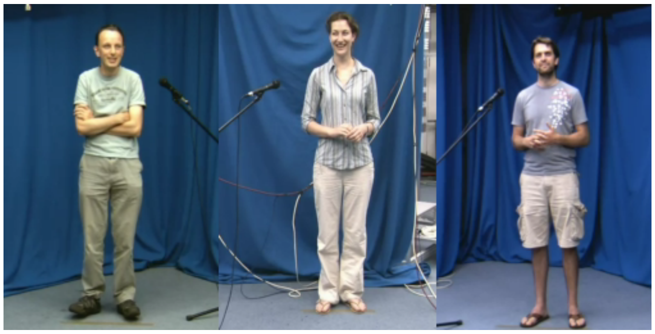

Fig. 1: Image showing full-body view of recorded video data of three people having a conversation.

4 Human Video Texture Synthesis

4.1 Video Data

The data set consists of approximately 30 minutes of video and audio recording of the full-body frontal view (516×340, 25 frames per second, 48kHz) and the close-up frontal face view (720×576, 25 frames per second, 48kHz) of 3 individuals having a conversation with each other. Each person remained in a stationary position relative to the camera as shown in Figure 1, although they were not constrained to do so, so there exists considerable ambient motion in their pose. Only the full-body video sequence is used for synthesis. The close-up fontal face video was only recorded to assist with the semi-supervised social signal labelling, which will be discussed in more detail later.

Each full-body video consists of approximately 43000 frames. To reduce computation complexity, the videos were reduced to grayscale and resized to a quarter of their original size. Given a video sequence $\mathbf{X}$, each frame is represented as a vector $\mathbf{x}_i$ where $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_{N_T}\}$ and N_T is the number of frames and $\mathbf{x}_i = (x_{i1}, y_{i1}, ..., x_{ix}, y_{iy}) \in \Re^{xy}$ for an x × y image.

To further reduce the complexity, Principal Component Analysis (PCA) is used for dimensionality reduction. The dimension of the feature space $|x_i|$ is reduced by projecting into the eigenspace d, where d is the chosen lower dimension $d \leq |x_i|$ such that $\sum_{i=1}^{d} \frac{\lambda_i}{\Sigma \forall \lambda} \geq .98$ or 98% of the energy is retained. $\mathbf{Y}$ is defined as a set of all points in the dimensionally reduced data where $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_{N_T}\}$ and $\mathbf{y}_i \in \Re^d$.

4.2 Identifying Cut Points

Similar to most work on motion synthesis, the motion data needs to be analysed to compute some measure of similarity between frames and derive points of intersection within the data. These points will allow transitions, either by 'motion blending' or 'switching', between different subsequences, producing motion paths not inherent in the original data.

The common approach is to compute the L2 distance over a window of frames in time and use a user defined threshold to balance between the quality of the transitions and the number of candidate transition points [1] [19] [18] [14] [15]. This approach works well, however, for very large data sets, it can be tedious to compute the distance between every frame. In our case, we would have to compute 3 separate 43000 × 43000 similarity matrices (one for each person), which would be time consuming and computationally expensive. Balci et al [16] proposed an iterative clustering procedure based on k-means to define clusters of poses suitable for transitions. However, k-means produces cluster centres not embedded in the data which can result in noise and outliers.

Instead, we adopt a k-medoid cluster algorithm to define N_C k-medoid points, where $N_C < N_T$. Each k-medoid point is defined as the local median in regions of high density, and can be used to define regions where appropriate transitions are possible. By only computing the L2 distance at these points, we reduce the amount of computation required to define candidate transitions, focusing attention on regions where transitions are most likely.

We define each k-medoid point as $\boldsymbol{\delta}_k \in \mathbf{Y}$. To preserve dynamics and account for temporal shape similarity, we compute a linearly weighted average of similarity over a fixed window of 0.25 seconds, centred on the k-medoids. Using a user defined threshold θ, the nearest points to each k-medoid points are identified to form clusters of cut points. The set containing the cut points of the n^{th} cluster is defined as $\mathbf{Y}_n^c = \{\mathbf{y}_{n,1}^c, ..., \mathbf{y}_{n,Q_n}^c\}$, where the number of cut points of the n^{th} cluster is denoted as Q_n. N_C is empirically determined based on the number of candidate cut points versus the quality of transitions and the amount of computation. In this work $N_C = 165$, greatly reducing the number L2 distance calculations.

Formally, each cut point belongs to a single cluster where plausible transitions can only be made between group members. This approach works well in data sets with high connectivity, however, less so for human video data where connectivity is limited. To overcome this problem, we extended the approach to allow cut points to belong to more than one cluster, providing the cut point clusters with more opportunities to perform novel movement.

For simplicity we define the transitions from the n^{th} cluster contents in the form $\{(\mathbf{y}_{n,1}^c, \mathbf{z}_{n,1}^c, \iota_{n,1}, C_{n,1}^z), ..., (\mathbf{y}_{n,Q_n}^c, \mathbf{z}_{n,Q_n}^c, \iota_{n,Q_n}, C_{n,Q_n}^z)\}$, where $\mathbf{y}_n^c$ is a cut point in the n^{th} cluster acting as the start transition point, $\mathbf{z}_n^c$ is the end transition point denoting the end of the subsequence between $\mathbf{y}_n^c$ and $\mathbf{z}_n^c$ (where $\mathbf{z}_n^c \in \mathbf{Y}^c$ and $\mathbf{z}_n^c \neq \mathbf{y}_n^c$), ι_n is the frame number of $\mathbf{y}_n^c$ in the original data, and C_n^z is the index of the cluster $\mathbf{z}_n^c$ belongs to.

4.3 Texture Motion Graph

So far, the cut points can be used to transition between different subsequences available in the data, however, there is no consideration to whether the transitions can perform all the available motion types or whether it leads to a dead-end. As a result, we pre-compute a *Texture Motion Graph* to guarantee global connectivity to different types of movements in the video data set.

Motion graph, proposed by Kovar et al [1], essentially connects various subsequences together to a form a directed graph, whereby the edges are the generated cut points. By assembling the graph, we can identify and eliminate cut points with low connectivity, improving reliability in the sequence selection process. Various forms of motion graph have been proposed in recent years [16] [15] [14] [22], built for animating motion captured data. *Texture Motion Graph* is specifically tailored for video textured data and extended to overcome ambiguities in transitioning between video frames.

Generating smooth blends between human video textures is a very challenging topic, since as human beings, we can easily recognise unnatural human movement or textures. Our similarity measure performs well with distinguishing different body poses, however does not account for facial gestures like laughing, talking and subtle changes in gaze direction. To overcome this, we use the social signal labels to assist the similarity measure. We prune the clusters of cut points to have the same gaze, talking, and laughing labels, as those of its k-medoid cluster centre, discarding cut points which do not match. This reduces the occurrences of rapid and unnatural changes in facial expressions out of context to the social interaction.

We define n strongly connected subgraphs for each unique set of social signals. This structure efficiently populates the graph with various links to social behaviour, making them easily accessible from any subgraph pose configuration. Social signals, such as head nods and head shakes, can occur in short quick bursts, lasting only a few seconds. By making available varying occurrences of a set of labels, we increase the opportunity of transitioning to a social behaviour quickly, and easily, making the graph more responsive. The Tarjan algorithm is used to derive the strongly connected subgraphs, and in our experiments we found $n = 3$ sufficient to populate our graph with varying social behaviour.

4.4 Dynamic Model

Conventional motion graph synthesis traverses the graph, connecting motion segments based on user specified constraints such as position, orientation and timing [1] [15]. Little interest is given to how common or likely the connecting nodes are given the data set. In smaller data sets, the user has limited choices of nodes to traverse, hence the quality of the chosen node of traversal is of little importance. However, in a densely populated data set, better quality transitions can be produced by computing the likelihood of a pose or frame as an additional parameterised weight. Hence, a dynamic model is learnt to derive the likelihood of pose in eigenspace based on the data set of video frames.

A statistical model of the constraints and dynamics present within the data can be created using a Probability Density Function (PDF). An appearance PDF is created using kernel estimation where each kernel $p(\mathbf{y}_i)$ is a Gaussian centred on a data example $p(\mathbf{y}_i) = G(\mathbf{y}_i, \Sigma)$. The likelihood of a posture or pose in eigenspace is modelled as a mixture of Gaussians using a multivariate normal distribution.

$$P(\mathbf{y}) = \frac{1}{N_T} \sum_{i=1}^{N_T} p(\mathbf{y}_i) \tag{1}$$

where the covariance of the Gaussian is:

$$\Sigma = \alpha \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_d} \end{pmatrix} \tag{2}$$

The width of the Gaussian in the i^{th} dimension is set to $\alpha\sqrt{\lambda_i}$. For all experiments $\alpha = 0.25$.

To reduce estimation time without sacrificing accuracy, a kd-tree is used to localise queries to neighbouring kernels, assuming the kernel estimation outside a local region contributes nominally to the local density estimation. Equation 1 is simplified to:

$$P'(\mathbf{y}) = \frac{1}{|\mathbf{Y}'|} \sum_{\forall \mathbf{y}_i \in \mathbf{Y}'} p(\mathbf{y}_i) \tag{3}$$

where $\mathbf{Y}' \subseteq \mathbf{Y}$, and $\mathbf{Y}'$ is a set containing the nearest neighbour kernels to $\mathbf{y}_i$ found efficiently with the kd-tree.

By learning a PDF, the data is represented in a generalised form which is analogous to a generative model. Using this form on its own, it is possible to generate novel motion frames, using pre-computed motion derivatives for a global approximation, combined with a gradient decent for optimisation. However, such a model runs the risk of smoothing out subtle motion details, and is not suitable for video. Instead, we combine the PDF with a Markov Transition Matrix to determine the likelihood of transitioning between cut points. This allows motion generation based on the original data, retaining important motion information.

Since transitions from one state to the next are dependent only on the current state, we define the conditional probability of moving from one cluster to another as $P(C_t|C_{t-1}) = p_{C_{t-1},C_t}$ where C_t is defined as the index for a cluster at time t.

4.5 Motion Synthesis

To generate novel motion sequences, the procedure is:

1. Given the current position in eigenspace $\mathbf{y}^c_t$, find all adjacent cut point neighbours in $\mathbf{Y}^c_t$ as defined in Section 4.2, to represent start transition points.
2. Find all associated end transition points $\mathbf{z}^c_{t,m}|m = \{1, ..., Q_t\}$. This gives a set of Q_t possible transitions from the starting point $\mathbf{y}^c_t$ in eigenspace.
3. Denote the cut point group index that $\mathbf{y}^c_t$ belongs to as C_t.
4. Calculate the likelihood of each transition as:

$$\boldsymbol{\phi}_m = P(C^z_{C_t,m}|C_t)P'(\mathbf{z}^c_{C_t,m}) \tag{4}$$

   where $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, .., \boldsymbol{\phi}_{Q_t}\}$.
5. Normalise the likelihoods such that $\sum_{i=1}^{Q_t} \boldsymbol{\phi}_i = 1$.

6. Since a maximum likelihood approach will result in repetitive animations, we randomly select a new start transition point $\mathbf{y}^c_{t,k}$ from $\boldsymbol{\Phi}$ based upon its likelihood as:

$$\arg\min_k \left(\sum_{j=1}^{k} \boldsymbol{\phi}_j \geq r \right) \tag{5}$$

   where k is the index of the newly chosen end transition point, $k \in m$, and r is a random number between 0 and 1, $r \in [(0, 1)]$.
7. All frames associated to the transition sequence between $\mathbf{y}^c_{t,k}$ and $\mathbf{z}^c_{t,k}$ are rendered.
8. The process then repeats from step (1) where $\mathbf{y}^c_{t+1} = \mathbf{z}^c_{t,k}$.

5 Social Interaction Model

5.1 Social Behaviour Experiment

Our data set [1] consists of recordings of 3 people in a conversation. We refer to the three individuals as person A, B, and C. Prior to capture, each person was given a questionnaire and asked to score from $1 - 3$ their general interest on a given set of book genres, film genres, and music genres. They were also given specific questions like: favourite sports, language(s) they spoke fluently, favourite music concerts, favourite theatre show etc. Their questionnaires were analysed to define 4 generic scenarios:

1. All interested in topic
2. Two people interested in topic, one person is not
3. One person interested in topic, two people are not
4. None are interested in topic

These 4 generic scenarios where derived from 8 topics of conversation as detailed in Table 1. The sixth column of Table 1 shows the limited duration of each topic, chosen to suite the scenario. A projector displayed the topic of conversation for discussion, and a quiet bell would ring to make the subjects aware of the change in topic. The subjects were unaware of the nature of the experiment, and were simple asked to discuss the topic displayed on the screen.

The aim of this experiment was to observe the social dynamics between the three people in scenarios when *interested* or *not interested* in the topics. To achieve this, we need to quantise their social behaviour in some form in order to obtain a clear distinction in social behaviour.

5.2 Semi-Supervised Social Signal Labelling

Pentland [23] proposed measuring non-linguistic social signals using four main observations: *activity level*, *engagement*, *emphasis* and *mirroring*. Using this as our base, we chose to observe 7 social signals in the conversation: *Voiced*, *Talking*, *Laughing*, *Head Shake*, *Head Nod*, *Activity Measure*, and *Gaze Direction*.

We use a variety of techniques to derive each label.

[1] The data set along with annotation can be made available upon request. Please email {d.okwechime@surrey.ac.uk}.

Scenario	A	B	C	Topic	Duration
1	3	3	3	Classical Music	5 minutes
2	3	3	1	Adventure Novels	5 minutes
3	3	1	3	Philosophy Novels	5 minutes
4	1	3	3	Rock Music	5 minutes
5	3	1	1	Sailing (Spoken in French)	2.5 minutes
6	1	3	1	Triathlon/Les Miserables (Spoken in Afrikaans)	2.5 minutes
7	1	1	3	Radio Head Concert	2.5 minutes
8	1	1	1	Horror Novels	1.5 minute

Table 1: Table showing 8 different social scenarios dictated by the topic of conversation. The three people are referred to as person **A**, **B**, **C**. The numbers indicate their interest in the topics where 3 is a high interest and 1 is a low interest

1. **Voiced[V]:** Audio stream represented using 12 MFCCs (Mel-Frequency Cepstral Coefficients) and a single energy feature of the standard HTK setup [24]. For each person, a few voiced segments were labelled and a Mahalanobis distance measure was used to derive a correlation between the voiced and non-voiced regions.
2. **Talking[T]:** With the voiced segments labelled, it was a simple process of labelling the voiced segments which were talking. This was done by hand.
3. **Laughing[L]:** The Viola-Jones face detector [25] was used to segment the face region in each frame. The lip region was localised by cropping the lower-centre region of the face. An AdaBoost classifier was then trained for laughing and used to label the remaining data.
4. **Head Shake[S]:** The Viola-Jones face detector was used to determine the movement of the face. Fast Fourier transform (FFT) was used to define high frequency movement along the x-axis
5. **Nod[N]:** Similar to head shakes, FFT was used to define high frequency movement along the y-axis.
6. **Activity Measure[A]:** The torso region of the full body video was segmented using colour and the mean-scaled standard deviation of velocity was measured. The leg and head regions are ignored because, there was no leg movement (subjects are stationary), and since we are more interested in gesture activity, changes in head posture/gaze would bias the activity measure.
7. **Gaze Direction[G]:** The eye pupils and the corners of the eyes were tracked using a Linear Predictor tracker [26]. The corners of the eyes were normalised to 0 and 1, and the position of the eye pupil within this region was used to determine if the person was gazing left [GL], right [GR] or centre [GC].

This produces N_T sets of social signal labels of 27 dimensions, where $1-9$ is for person A, $10-18$ for person B and $19-27$ for person C. We define 2 complete sets of social signal vectors for *interested* and *not interested* scenarios as $\mathcal{F}_{(Int)}$, and $\mathcal{F}_{(NoInt)}$ such that $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^{N_T}$ where $\mathbf{f}_i$ is a 27 dimensional binary vector.

5.3 Data Mining for Social Trends

This framework is driven by the speaker. At any given time, there is only one speaker and two listeners. We are interested in the combination of social signals a listener performs given a speaker's social behaviour when the listener is *interested* and *not interested* in the conversation. Manually observing all combinations of listener and speaker behaviours in such a large data set would be virtually impossible. A solution would be to make some common sense prior assumptions of expected trends (i.e. an *interested* listener would gaze more at the speaker than when they are *not interested*) and focus primarily on these assumptions. However, there is no way of proving or disproving such assumptions, and, there is a large list to chose from.

We propose a novel approach to deriving social dynamics and trends between the subjects based on data mining [2]. Data mining allows for large data sets to be mined to identify the reoccurring patterns within the data in an efficient manner. In this framework, Apriori Association rule [2] [27] mining is used. Formally developed for supermarkets to analyse millions of customer's shopping trends, we aim to find *association rules* within the numerous combinations of social trends between the subjects in an *interested* and *not interested* scenario given the speaker's social behaviour.

An association rule is a relationship of the form $\{R_i^A\} \Rightarrow R_i^C$ where R_i^A is a set of social signals of the speaker, and R_i^C a sets of social signals of the listener. $R_i^A = \{r_{i,1}^A, ..., r_{i,|R_i^A|}^A\}$ is the antecedent where r_i^A denotes a speaker's social signal, and $R_i^C = \{r_{i,1}^C, ..., r_{i,|R_i^C|}^C\}$ the consequence where r_i^C is a listener's social signal. An example would be, if $R_1^A = \{[T], [N]\}$, and $R_1^C = \{[N]\}$ as defined in Section 5.2, then, $\{R_1^A\} \Rightarrow R_1^C$ would imply 'when person A is talking and nods, person B is very likely to also nod'. The belief of each rule is measured by a *support* and *confidence* value. The *support* measures the statistical significance of a rule, it is the probability that a transaction contains itemset R_i^A.

$$sup(\{R_i^A\} \Rightarrow R_i^C) = sup(\{R_i^A\} \cup R_i^C) \quad (6)$$

The *confidence* is the number of occurrences in which the rule is correct, relative to the number of cases in which it is applicable.

$$conf = \frac{sup(\{R_i^A\} \cup R_i^C)}{sup(R_i^A)} * 100 \quad (7)$$

Apriori Association mining is applied to the social signal labels for both *interested* listener and *not interested* listener scenarios, to derive frequently occurring association rules. We define the set of all rules extracted using data mining as:

$$R = \{(R_i^A \Rightarrow R_i^C, conf_i)\}_{i=1}^{|R|} \quad (8)$$

where the total number of rules is $|R|$.

Traditionally, data mining looks for a combination of symbols that occur simultaneously, however, a listener's social behaviour is always a response to the speaker's social signals, hence, co-articuation is not possible. To account for this, *temporal bagging* within a set temporal window is used to enforce a temporal coherence between features. Given a speaker's social signal, we observe the listener's social behaviour $s = 10$ frames in the future (approx $\frac{1}{2}$ a second).

6 Social Interactive Motion Control using Apriori Mining

Since we are only interested in deriving animations of the listener, we compute the conditional probability of the listener's social response given the speaker's social signals, as weighted variables to control the motion model.

Given the chosen speaker's 9 dimensional binary vector $\mathbf{f}_t$ (as explained in Section 5.2) at time t, where $\mathbf{f}_t \subset \mathbf{f}_\iota$ (ι is the frame index of the current query cut point as detailed in Section 4.2), we derive the power set $2^{\mathbf{f}_t}$ for all combinations of the speaker's active social signals. We find a suitable matching set of rules $R^t \subset R$ such that $\forall (R_j^{A,t} \Rightarrow R_j^{C,t}, conf_j) \in R^t$, where there exists $\mathbf{f} \in 2^t$ and $\mathbf{f} \subset R_j^A$. The weighted combination of the results are obtained as follows:

$$W = \sum_{j=1}^{|R^t|} conf_j I(R_j^{C,t}, \mathbf{f}_\iota) \tag{9}$$

where

$$I(R_j^C, \mathbf{f}) = \begin{cases} 1 \text{ if } \mathbf{f} \subset R_j^C \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

In the motion synthesis process, Equations 4 is altered as follows, for all:

$$\boldsymbol{\phi}_m = P(C^z_{C_t,m}|C_t).P'(\mathbf{z}^c_{C_t,m}).W \tag{11}$$

7 Animation/Results

To validate our experiment, we observe the trends in the mined *confidence* values, which details the likelihood of a listener's social response given a speaker's social signal. 1350 rules were extracted from the mining in the *interested* scenario, and 1400 in the *not interested* scenario, which resulted in 1034 matching rules in both scenarios. Dividing the *confidence* values in the *interested* scenario by those in the *not interested* scenario for matching rules, we obtain results of greater than 1 when the rules occur more frequently in the *interested* scenario, and less than 1 when they occur more in the *not interested* scenario. The results are shown in Table 2. We are unable to show all combinations of association rules so we show the highest trends greater than 10.

The association rule labels are as detailed in Section 5.2. To add clarity to the gaze labels, instead of [GL], [GR] and [GC], we use [GA], [GB], [GC], [GN], representing *gazing at person A*, *gazing at person B*, *gazing at person C*, and *gazing at no one*, respectively. This allows us to know who the speaker gazes at.

Rows 2 and 3 in Table 2 present the most prominent trends. Row 2 suggests that 'when person B is talking and gazing at no one, person A talks'. In

No.	**Association Rules**	$\frac{\mathbf{cv(int)}}{\mathbf{cv(not)}}$	No.	**Association Rules**	$\frac{\mathbf{cv(int)}}{\mathbf{cv(not)}}$
1	$\{C = [S] \Rightarrow A = [\mathbf{T}]\}$	11	11	$\{B = [N] \Rightarrow C = [\mathbf{T}]\}$	13
2	$\{B = [GN] \Rightarrow A = [\mathbf{T}]\}$	48	12	$\{B = [GN] \Rightarrow C = [\mathbf{T}]\}$	11.6
3	$\{C = [GA] + [S] \Rightarrow A = [\mathbf{V}]\}$	74	13	$\{B = [A] \Rightarrow C = [\mathbf{T}]\}$	11
4	$\{C = [GN] + [N] \Rightarrow A = [\mathbf{N}]\}$	40.2	14	$\{B = [GC] \Rightarrow A = [\mathbf{T}]\}$	18
5	$\{C = [A] + [N] \Rightarrow A = [\mathbf{N}]\}$	12	15	$\{A = [L] + [S] \Rightarrow B = [\mathbf{T}]\}$	11.9
6	$\{A = [GN] + [A] \Rightarrow B = [\mathbf{A}]\}$	38.5	16	$\{B = [GC] + [N] \Rightarrow C = [\mathbf{T}]\}$	13.3
7	$\{A = [L] + [GB] \Rightarrow C = [\mathbf{T}]\}$	33	17	$\{B = [A] + [N] \Rightarrow C = [\mathbf{L}]\}$	13.4
8	$\{B = [GC] + [A] \Rightarrow C = [\mathbf{T}]\}$	25.3	18	$\{C = [GN] + [S] \Rightarrow A = [\mathbf{N}]\}$	11
9	$\{B = [GC] + [A] \Rightarrow C = [\mathbf{N}]\}$	46	19	$\{A = [GN] + [N] \Rightarrow C = [\mathbf{T}]\}$	11
10	$\{C = [GN] \Rightarrow A = [\mathbf{S}]\}$	10	20	$\{C = [L] + [S] \Rightarrow A = [\mathbf{N}]\}$	12

Table 2: Table showing the highest trends ($approx > 10$) for a set of rules for confidence values from the *interested* scenarios cv(int) divided by the confidence values for the *not interested* scenario cv(not), for the individual people. Association rule labels are as detailed in Section 5.2

this context, person A talking suggests turn-taking, showing more interest in participating in the conversation in an *interested* scenario as opposed to a *not interested* scenario. Row 3 suggest that 'when person C is talking, gazing at person A and shaking their head, person A is voiced'. Voiced regions imply an exchange of short single words like 'uh-huh' or 'yea', used by a listener to express acknowledgement and understand to the speaker. Other high trends like in rows 4, 5, and 6 also suggests mirroring, where the listener mimics the speaker's social signal such as *nods*, and *active* body movement. Looking at these general contrasts, we see that *talking* is a highly common response from an *interested* listeners, so we can safely assume that turn-taking is an important measure of social interest.

These quantitative results prove there is a clear distinction between an *interested* and *not interested* listener in a social context, and our social experiment provides an appropriate means of modelling levels of social interest.

Using the motion model, the user is given control over the various combinations of social behaviour of the human video avatars, however, not all combinations of social behaviour controls are possible. This is the case for all avatars with regards to performing *head shakes*. This is mostly due to the limited availability of the particular social behaviour in the data set, resulting in very limited connectivity in the texture motion graph. Regardless, most of the popular combinations of social behaviour like *laughing* and *head nods* are connected and responsive.

Not only can the user control the social behaviour of the video avatars but, using the data mined *confidence* values, autonomous interaction is possible. As shown in Figure 2, the user has interactive control over who speaks, and the level of interests of the listeners. With these set parameters, the avatars interact appropriate, traversing the texture motion graph to attach video subsequences together, guided by the data mined conditional weights. Figure 2 (D), further demonstrate the approach by producing autonomous interaction generated by identical avatars.

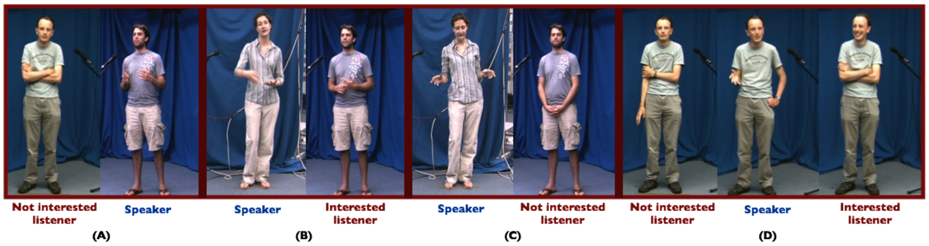

Fig. 2: Image showing human video texture synthesis of conversational avatars. (A), (B) and (C) show different generated videos given the scenarios of an *interest* and *not interested* listener in the conversation. (D) demonstrates the diversity of the approach, allowing identical avatars to socially interact together.

8 Conclusion

Our social dynamics model is able to derive trends between a speaker and listener in a conversation. We successfully parameterise these trends using data mining to derive the conditional probability of a listener's behaviour given a speaker's social signal. Human video motion modelling using a texture motion graph produces plausible transitions between cut points, allowing interactive control over a video avatar's social behaviour. Future work will include video blending strategies for photorealistic data to fix glitches occurring on the seams between video segments.

Utilising the social dynamics model to drive the animation, the user can alter the interest level of participants in the conversation, effectively changing their social responses. This approach can be extended to other social scenarios such as *conflicts*, and other synthesis approaches such as *motion captured* animations, and future work will extend the observed social signals within the system.

Acknowledgements. This work was supported by the EPSRC project LILiR (EP/E027946/1) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 231135 - DictaSign.

References

1. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In Proc. of ACM SIGGRAPH, 21, 3, Jul (2002) 473–482
2. Agrawal, A., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: In Proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data SIGMOD'93. (1993)
3. Szummer, M., Picard, R.: Temporal texture modeling. In: In Proc. of IEEE Int. Conf. on Image Processing, 1996. (1996) 823–826
4. Efros, A., Leung, T.: Texture synthesis by non-paramteric sampling. In: In Int. Conf. on Computer Vision. (1999) 1033–1038
5. Kwatra, V., Schodl, A., Essa, I., Turk, G., Bobick, A.: Graphcut textures. In: ACM Trans. on Graphics, SIGGRAPH 2003, 22, 3. (2003) 277–286
6. Bhat, K., Seitz, S., Hodgins, J., Khosla, P.: Flow-based video synthesis and editing. In: ACM Trans. on Graphics, SIGGRAPH 2004. (2004)
7. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. J. Vis. **2** (2002) 371–387

8. Pullen, K., Bregler, C.: Synthesis of cyclic motions with texture (2002)
9. Okwechime, D., Bowden, R.: A generative model for motion synthesis and blending using probability density estimation. In: Fifth Conference on Articulated Motion and Deformable Objects, 9-11 July, Mallorca, Spain (2008)
10. Tanco, L.M., Hilton, A.: Realistic synthesis of novel human movements from a database of motion captured examples. In Proc. of the IEE Workshop on Human Motion HUMO 2000) (2000)
11. Arikan, O., Forsyth, D., O'Brien, J.: Motion synthesis from annotation. In ACM Transaction on Graphics, 22, 3, July, (SIGGRAPH 2003) (2003) 402–408
12. Okwechime, D., Ong, E.J., Bowden, R.: Real-time motion control using pose space probability density estimation. In: IEE Int. Workshop on Human-Computer Interaction. (2009)
13. Treuille, A., Lee, Y., Popovic, Z.: Near-optimal character animation with continuous control. In: Proceedings of SIGGRAPH 2007 26(3). (2007)
14. Rachel, H., Gleicher, M.: Parametric motion graph. 24th Int. Symposium on Interactive 3D Graphics and Games (2007) 129–136
15. Shin, H., Oh, H.: Fat graphs: Constructing an interactive character with continuous controls. In: Proc. of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation. (2006) 298
16. Balci, K., Akarun, L.: Generating motion graphs from clusters of individual poses. 24th Int. Symposium on Computer and Information Sciences (2009) 436–441
17. Lee, J., Chai, J., Reitsma, P., Hodgins, J., Pollard, N.: Interactive control of avatars animated with human motion data. ACM Trans. on Graphics **21** (2002) 491–500
18. Schödl, A., Szeliski, R., Salesin, D., Essa, I.: Video textures. In: Proc. of the 27th annual conf. on Computer graphics and interactive techniques, SIGGRAPH 2000, ACM Press/Addison-Wesley Publishing Co. New York (2000) 489–498
19. Flagg, M., Nakazawa, A., Zhang, Q., Kang, S., Ryu, Y., Essa, I., Rehg, J.: Human video textures. In: Proc. of the 2009 symposium on Interactive 3D graphics and games, ACM (2009) 199–206
20. Ekman, P., Friesen, W.: Facial action coding system. In: Consulting Psychologists Press, Palo Alto, CA. (1977)
21. Argyle, M.: Bodily communication. In: Methuen. (1987)
22. Beaudoin, P., Coros, S., van de Panne, M., Poulin, P.: Motion-motif graphs. In: In Proc. of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. (2008) 117–126
23. Pentland, A.: A computational model of social signaling. In: 18th Int. Conf. on Pattern Recognition. ICPR. (2006)
24. Mertins, A., Rademacher, J.: Frequency-warping invariant features for automatic speech recognition. In: 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. Volume 5. (2006)
25. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Proc. IEEE CVPR 2001. (2002)
26. Ong, E.J., Lan, Y., Theobald, B.J., Harvey, R., Bowden, R.: Robust facial feature tracking using selected multi-resolution linear predictors. In: Int. Conf. Computer Vision ICCV 2009. (2009)
27. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: In VLDB'94 Proc. of 20th Int. Conf. on Very Large Data Bases. (1994) 487–499