

Conditional Sequence Model for Context-based Recognition of Gaze Aversion

Louis-Philippe Morency and Trevor Darrell

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139
{lmorency, trevor}@csail.mit.edu

Abstract. Eye gaze and gesture form key conversational grounding cues that are used extensively in face-to-face interaction among people. To accurately recognize visual feedback during interaction, people often use contextual knowledge from previous and current events to anticipate when feedback is most likely to occur. In this paper, we investigate how dialog context from an embodied conversational agent (ECA) can improve visual recognition of eye gestures. We propose a new framework for contextual recognition based on Latent-Dynamic Conditional Random Field (LDCRF) models to learn the sub-structure and external dynamics of contextual cues. Our experiments show that adding contextual information improves visual recognition of eye gestures and demonstrate that the LDCRF model for context-based recognition of gaze aversion gestures outperforms Support Vector Machines, Hidden Markov Models, and Conditional Random Fields.

Key words: Contextual information, Conditional Random Fields, Eye gesture recognition, gaze aversion

1 Introduction

In face to face interaction, eye gaze is known to be an important aspect of discourse and turn-taking. To create effective conversational human-computer interfaces, it is desirable to have computers which can sense a user’s gaze and infer appropriate conversational cues. Embodied conversational agents, either in robotic form or implemented as virtual avatars, have the ability to demonstrate conversational gestures through eye gaze and body gesture, and should also be able to perceive similar displays as expressed by a human user.

Previous work has shown that human participants avert their gaze (i.e. perform “look-away” or “thinking” gestures) to hold the conversational floor even while answering relatively simple questions [1]. A gaze aversion gesture while a person is thinking may indicate that the person is not finished with their conversational turn. If the ECA senses the aversion gesture, it can correctly wait for mutual gaze to be re-established before taking its turn.

When recognizing visual feedback, people use more than their visual perception. Knowledge about the current topic and expectations from previous utterances help guide our visual perception in recognizing nonverbal cues. Context

information can be found from cues like the words and prosody/punctuation (e.g., word pair “do you” with question mark) of the current sentence but the real meaning and structure of these cues can sometimes be hidden (e.g., this is a yes/no question). The dynamic between these contextual cues (e.g., “do you” before the question mark) is also relevant information. An important challenge for context-based recognition is to learn these hidden sub-structures and external dynamics from the contextual cues.

In this paper, we present a framework for context-based recognition that uses Latent-Dynamic Conditional Random Field (LDCRF) models [2] to learn the hidden sub-structure and external dynamic of contextual information. The main two contributions of this paper are that we are the first to (1) show that dialog context can improve gaze aversion recognition and (2) demonstrate that LDCRF models are superior to other learning methods (i.e., SVM, CRF, and HMM) at learning relevant context and integrating it with visual observations for gaze aversion recognition. The power of LDCRFs comes from the fact that it learns the extrinsic dynamics by modeling a continuous stream of class labels, and learns internal sub-structure by utilizing intermediate hidden states.

The remainder of this paper is organized as follows. In Section 2 we review relevant related work, and in Section 3 we present our LDCRF context-based recognition framework. The details of our three set of experiments including information about the dataset, the compared models and the methodology are described in Section 4. We present and discuss the results of our experiments in Section 5. Finally, a summary and discussion of future work are provided in Section 6.

2 Related Work

Eye gaze plays an important role in face-to-face interactions. Kendon proposed that eye gaze in two-person conversation offers different functions: monitoring visual feedback, expressing emotion and information, regulating the flow of the conversation (turn-taking), and improving concentration by restricting visual input [3]. Many of these functions have been studied to create more realistic ECAs [4-6], but they have tended to explore only gaze directed towards individual conversational partners or objects.

A considerable body of work has been carried out regarding eye gaze and eye motion patterns for perceptive user interfaces. Velichkovsky suggested the use of eye motion to replace the mouse as a pointing device [7]. Qvarfordt and Zhai used eye-gaze patterns to sense the user interest with a map-based interactive system [8]. Li and Selker developed the InVision system which responded to a user’s eye fixation patterns in a kitchen environment [9].

Context has been previously used in computer vision to disambiguate recognition of individual objects given the current overall scene category [10]. Fujie *et al.* also used HMMs to perform head nod recognition [11]. In their paper, they combined head gesture detection with prosodic low-level features computed from

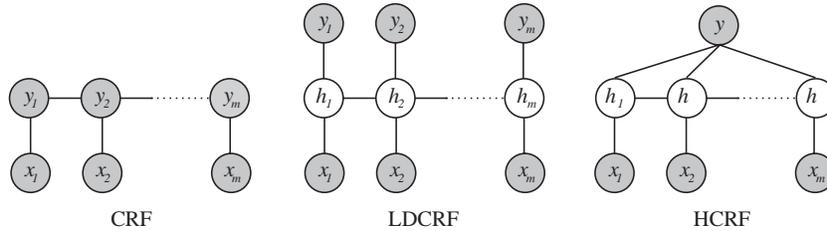


Fig. 1. Comparison of the LDCRF model [2] with two related models: CRF [12] and HCRF [14, 15]. In these graphical models, x_j represents the j^{th} observation (corresponding to the j^{th} frame of the video sequence), h_j is a hidden state assigned to x_j , and y_j the class label of x_j (i.e. head-nod or other-gesture). Gray circles are observed variables. The LDCRF model combines the strengths of CRFs and HCRFs in that it captures both extrinsic dynamics and intrinsic structure and can be naturally applied to predict labels over unsegmented sequences.

Japanese spoken utterances to determine strongly positive, weak positive and negative responses to yes/no type utterances.

The use of dialogue context for visual gesture recognition was first explored in [18]. In [18] they propose a late-fusion framework for incorporating dialog context in head gesture recognition. This framework was later extended to include context from conventional graphical user interfaces [?]. In both papers, the experiments were performed on head gesture recognition. This paper is the first to extend the idea of context-based recognition to recognize eye gesture. Also, the approach presented in [18, ?] used multi-class SVMs to train the context-based recognizer. Unlike LDCRFs, SVMs do not model the external dynamics between classes and do not explicitly model hidden sub-structure.

LDCRF models offer several advantages over previous discriminative models (see Figure 1). In contrast to Conditional Random Fields (CRFs) [12], our method incorporates hidden state variables which model the sub-structure of gesture sequences. The CRF approach models the transitions between gestures, thus capturing extrinsic dynamics, but lacks the ability to learn the internal sub-structure. In contrast to Hidden-state Conditional Random Fields (HCRFs) [13], our method can learn the dynamics between gesture labels and can be directly applied to label unsegmented sequences. The results reported in [2] demonstrate that LDCRF outperforms models based on Support Vector Machines (SVMs), HMMs, CRFs and HCRFs on visual gesture recognition task. In this paper, we demonstrate that LDCRF models are superior to other learning methods at learning relevant context and integrating it with visual observations.

3 Context-based Recognition Framework using LDCRF

For reliable recognition of visual feedback during face-to-face conversational interactions, people use knowledge about the current dialogue to anticipate gestures from their interlocutors.

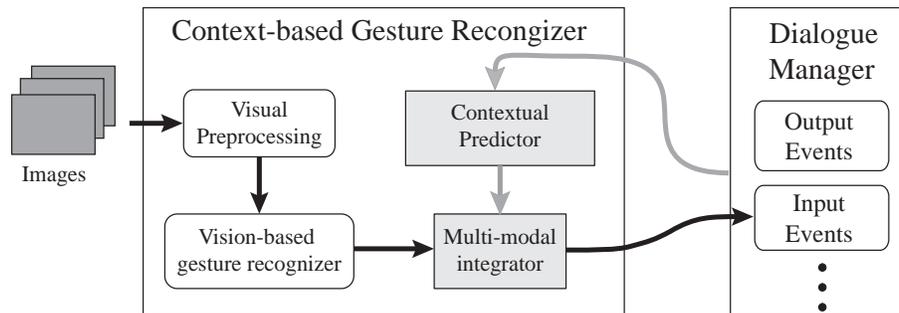


Fig. 2. Framework for context-based gesture recognition. The contextual predictor translates contextual features into a likelihood measure, similar to the visual recognizer output. The multi-modal integrator fuses these visual and contextual likelihood measures.

We can use a conversational agent’s knowledge about the current dialogue to improve recognition of visual feedback (i.e., eye gestures). The dialogue manager merges information from the input devices with the history and the discourse model [16, 17]. The dialogue manager contains two main sub-components, an agenda and a history: the agenda keeps a list of all the possible actions the agent and the user (i.e., human participant) can do next. This list is updated by the dialogue manager based on its discourse model (prior knowledge) and on the history. Dialogue managers generally exploit contextual information to produce output for the speech and gesture synthesizer, and we can use similar cues to predict when user visual feedback is most likely.

Following [18], we use three types of contextual features easily available from the dialogue manager: lexical features, prosody and punctuation, and timing. The contextual information is extracted from the dialogue manager rather than directly accessing internal ECA states. This strategy makes it possible to extract dialogue context without any knowledge of the internal representation and therefore makes it applicable to most ECA architectures. Figure 2 shows the general architecture of the context-based recognition framework.

In the following subsections we first give a formal description of the LD-CRF and then show how LDCRF is integrated in the context-based recognition framework.

3.1 LDCRF Model

As described in [2], the task of the LDCRF model is to learn a mapping between a sequence of observations $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ and a sequence of labels $\mathbf{y} = \{y_1, y_2, \dots, y_m\}$. Each y_j is a class label for the j^{th} frame of a video sequence and is a member of a set \mathcal{Y} of possible class labels, for example, $\mathcal{Y} = \{\text{gaze-aversion}, \text{other-gesture}\}$. Each frame observation x_j is represented by a feature vector $\phi(x_j) \in \mathbf{R}^d$, for example, the head velocities at each

frame. For each sequence, we also assume a vector of “sub-structure” variables $\mathbf{h} = \{h_1, h_2, \dots, h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Given the above definitions, we define a latent conditional model:

$$P(\mathbf{y} | \mathbf{x}, \theta) = \sum_{\mathbf{h}} P(\mathbf{y} | \mathbf{h}, \mathbf{x}, \theta) P(\mathbf{h} | \mathbf{x}, \theta). \quad (1)$$

where θ are the parameters of the model.

To keep training and inference tractable, we restrict the LDCRF model to have disjoint sets of hidden states associated with each class label. Each h_j is a member of a set \mathcal{H}_{y_j} of possible hidden states for the class label y_j . We define \mathcal{H} , the set of all possible hidden states, to be the union of all \mathcal{H}_{y_j} sets. Since sequences which have any $h_j \notin \mathcal{H}_{y_j}$ will by definition have $P(\mathbf{y} | \mathbf{h}, \mathbf{x}, \theta) = 0$, we can express the LDCRF model as:

$$P(\mathbf{y} | \mathbf{x}, \theta) = \sum_{\mathbf{h}: \forall h_j \in \mathcal{H}_{y_j}} P(\mathbf{h} | \mathbf{x}, \theta). \quad (2)$$

Given a training set consisting of n labeled sequences $(\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1 \dots n$, training is done following [19, 12] using this objective function to learn the parameter θ^* :

$$L(\theta) = \sum_{i=1}^n \log P(\mathbf{y}_i | \mathbf{x}_i, \theta) - \frac{1}{2\sigma^2} \|\theta\|^2 \quad (3)$$

The first term in Eq. 3 is the conditional log-likelihood of the training data. The second term is the log of a Gaussian prior with variance σ^2 , i.e., $P(\theta) \sim \exp\left(-\frac{1}{2\sigma^2} \|\theta\|^2\right)$.

For testing, given a new test sequence \mathbf{x} , we want to estimate the most probable sequence labels \mathbf{y}^* that maximizes our conditional model:

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} \sum_{\mathbf{h}: \forall h_i \in \mathcal{H}_{y_i}} P(\mathbf{h} | \mathbf{x}, \theta^*) \quad (4)$$

For a more detailed discussion of LDCRF training and inference see [2].

3.2 LDCRF Context-based Recognition

The contextual predictor outputs a likelihood measurement at the same frame rate as the vision-based recognizer so that the multi-modal integrator can merge both measurements. For this reason, feature vectors \mathbf{x}_j are computed at every frame j (even though the contextual features do not directly depend on the input images).

For the LDCRF model, the likelihood measurement for a specific gesture g is equal to the marginal probability $P(y_j = g | \mathbf{x}, \theta^*)$. This probability is equal

to the sum of the marginal probabilities for the hidden states part of the subset \mathcal{H}_g :

$$P(y_j = g \mid \mathbf{x}, \theta^*) = \sum_{\mathbf{h}: \forall h_j \in \mathcal{H}_g} P(\mathbf{h} \mid \mathbf{x}, \theta^*) \quad (5)$$

where \mathbf{x} is the concatenation of all the feature vectors \mathbf{x}_j for the entire sequence and θ^* are the model parameters learned during training. When testing offline, the marginal probabilities are computed using a forward-backward belief propagation algorithm. To estimate the marginal probabilities online, it is possible to define \mathbf{x} as the concatenation of all feature vectors up to frame j and use the forward-only belief propagation algorithm.

Our integration component takes as input the likelihood measurement from the contextual predictor and the visual observations from the vision-based head gesture recognizer, and recognizes whether a head gesture has been expressed by the human participant. The output from the integrator is further sent to the dialogue manager or the window manager so it can be used to decide the next action of the ECA.

4 Experiments

We designed our experiments to demonstrate how contextual information can improve eye gesture recognition and to demonstrate the superior performance of LDCRF on context-based recognition compared to baseline methods. We performed three series of experiments:

Experiment 1 where we compare the vision-only approach with the context-based recognition using LDCRF models. The goal of this experiment is to show that dialog context can improve eye gesture recognition

Experiment 2 where we compare the LDCRF model to SVM, CRF and HMM models for context-based recognition of gaze aversion. In this set of experiments, the contextual predictor and the multimodal integrator are both trained using the same model (either LDCRF, SVM, CRF or HMM). The goal of this experiment is to show the superiority of LDCRF for context-based recognition.

Experiment 3 where we first train the contextual predictor with the LDCRF model and then train the multimodal integrator with one of the four model. The goal of this experiment is to analyze the relative importance of LDCRF for contextual prediction and multimodal integration.

In the following subsections, we first describe our dataset used in our experiments, then present the models used to compare the performance of the LDCRF model, and finally describe our experimental methodology.

4.1 Eye Gesture Dataset

Our dataset came from a user study that shown that human participants naturally perform gaze aversion gestures when interacting with an avatar [1]. The

goal of this dataset is to differentiate gaze aversion gestures from all other type of eye gestures (e.g., eye contact or deictic gestures). Our dataset consist of 6 human participants interacting with a virtual embodied agent. Each video sequence lasted approximately 10-12 minutes, and was recorded at 30 frames/sec, for a total of 105,743 frames. During these interactions, human participants would rotate their head up to +/-70 degrees around the Y axis and +/-20 degrees around the X axis, and would also occasionally move their head, mostly along the Z axis.

The dataset was labeled with the start and end points of each gaze aversion gestures. Each frame was labeled either as gaze-aversion or as other-gesture which included sections of video where people were looking at the avatar or performing deictic gestures. The contextual cues from the dialogue manager (spoken utterances with start time and duration) were recorded during each interaction and were later automatically processed to create the contextual features necessary for the contextual predictor. The previous section showed how the contextual features are automatically computed.

For each video sequence, the eye gaze was estimated using the view-based appearance model described in [1] and for each frame a 2-dimensional eye gaze estimate was obtained. The eye gaze estimates were logged online with the contextual cues. For this dataset, the vision-based recognizer is a LDCRF model trained and validated offline on the same training and validation sets used for the contextual predictor and the multi-modal integrator.

4.2 Models

In our experiments, the LDCRF model is compared with three models: Conditional Random Field (CRF), Hidden Markov Model (HMM) and Support Vector Machine (SVM).

Conditional Random Field As a first baseline, we trained a single CRF chain model where every gesture class has a corresponding state label. During evaluation, marginal probabilities were computed for each state label and each frame of the sequence using belief propagation. The optimal label for a specific frame is typically selected as the label with the highest marginal probability. In our case, to be able to plot ROC curves of our results, the marginal probability of the primary label (i.e. **gaze-aversion**) was thresholded at each frame, and the frame was given a positive label if the marginal probability was larger than the threshold. The objective function of the CRF model contains a regularization term similar to the regularization shown in Equation 3 for the LDCRF model. During training and validation, this regularization term was validated with values 10^k , $k = -3..3$.

Support Vector Machine As a second baseline, a multi-class SVM was trained with one label per gesture using a Radial Basis Function (RBF) kernel. Since the SVM does not encode the dynamics between frames, the training set was decomposed into frame-based samples, where the input to the SVM is the head

velocity or eye gaze at a specific frame. The output of the SVM is a margin for each class. This SVM margin measures how close a sample is to the SVM decision boundary [20]. The margin was used to plot the ROC curves. During training and validation, two parameters were validated: C , the penalty parameter of the error term in the SVM objective function, and γ , the RBF kernel parameter. Both parameters were validated with values $10^k, k = -3..3$.

Hidden Markov Model As a third baseline, an HMM was trained for each gesture class. We trained each HMM with segmented subsequences where the frames of each subsequence all belong to the same gesture class. This training set contained the same number of frames as the one used for training the other models except frames were grouped into subsequences according to their label. The HMMs trained on subsequences are concatenated into a single HMM with the number of hidden states equal to the sum of hidden states from each individual HMM. For example, if the recognition problem has two labels and each individual HMM is trained using 3 hidden states, then the concatenated HMM will have 6 hidden states. To estimate the transition matrix of the concatenated HMM, we compute the Viterbi path of each training subsequence, concatenate the subsequences into their original order, and then count the number of transitions between hidden states. The resulting transition matrix is then normalized so that its rows sum to one. At testing, we apply the forward-backward algorithm on the new sequence, and then sum at each frame the hidden states associated with each class label. The resulting HMM can be seen as a generative version of our LDCRF model. During training and validation, we varied the number of states from 1 to 6 and the number of Gaussians per mixture from 1 to 3.

Latent-Dynamic Conditional Random Field Our LDCRF model was trained using the objective function described in [2]. During evaluation, we compute ROC curves using the maximal marginal probabilities of Equation 4. During training and validation, we varied the number of hidden states per label (from 2 to 6 states per label) and the regularization term (with values $10^k, k = -3..3$).

4.3 Methodology

In our experiments, the vision-based recognizer was trained and tested using LDCRF since this model gave the best performance for the visual recognition task (see [2] for details). The contextual predictors and multi-modal integrator (also referred to as “Fusion” in the result section) were trained using one of the four models described in the previous subsection. The contextual features were computed from the dialog context of the avatar using the technique described in [18].

The experiments were performed using a leave-one-out testing approach. For validation, we did holdout cross-validation where a subject is randomly picked from the training set and kept for validation. The optimal validation parameters were picked based on the equal error rate for the validation set.

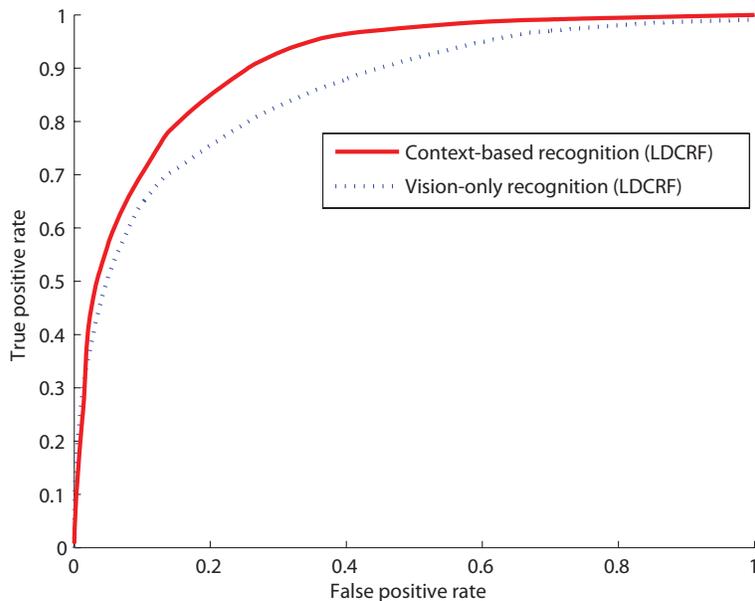


Fig. 3. Results from Experiment 1 comparing vision-only approach with context-based recognition using LDCRF models. We can see that dialog context significantly improves (p-value = 0.043) the gaze aversion recognition performance.

The dataset contained an unbalanced number of **other-gesture** frames. To have a balanced training set and reduce the training time, the training dataset was preprocessed to create a smaller training dataset containing an equal number of **other-gesture** and *transition* subsequences. Each *transition* subsequence includes frames from one complete **gesture** subsequence and frames before and after the gesture labeled as **other-gesture**. The size of the **other-gesture** gap before and after the gesture was randomly picked between 2 and 50 frames. The number of transition subsequences was equal to the number of ground truth gestures in the original training set. **Other-gesture** subsequences were randomly extracted from the original sequences with length varying between 30-60 frames.

5 Results and Discussion

For the ROC curves shown in this section, the true positive rate is computed by dividing the number of recognized frames by the total number of ground truth frames. Similarly, the false positive rate is computed by dividing the number of falsely recognized frames by the total number of **other-gesture** frames.

Figure 3 shows the results of Experiment 1 where we compare the LDCRF vision-only approach with the LDCRF context-based approach. We can see in this figure that context information does improve recognition of eye gesture. The ROC curve of LDCRF combining both vision and context is higher than that of

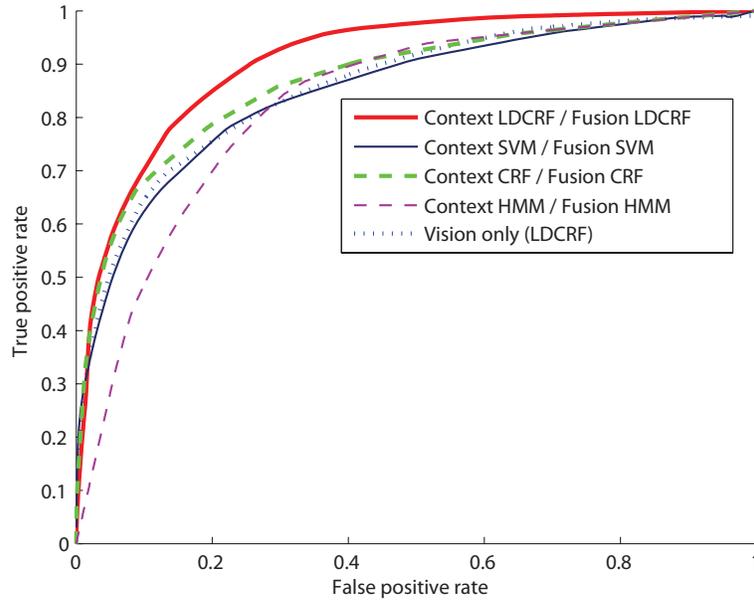


Fig. 4. Results from Experiment 2 comparing the LDCRF model to SVM, CRF and HMM models for context-based recognition of gaze aversion. Both the contextual predictor and the multimodal integrator were trained using the same model. The ROC curves show the performance of each trained multimodal integrator. LDCRF outperforms all three other models with statistically significant differences for SVM and HMM (p-values equal to 0.0329, and 0.0343 respectively).

LDCRF using only vision without context. Using t-test analysis, the difference between the two curves, calculated based on the equal error rates, is statistically significant (one-tail $p = 0.043$).

Figure 4 shows the results from Experiment 2 where we compare the LDCRF model to SVM, CRF and HMM models for context-based recognition of gaze aversion. LDCRF outperforms all three other models (SVM, CRF and HMM) for context-based recognition. A paired t-test analysis over all tested subjects returns a one-tail p-value of 0.0329, 0.0717 and 0.0343 when comparing the equal error rate performance of LDCRF with SVM, CRF and HMM respectively. This analysis shows statistically significant improvement using the LDCRF model when compared to both SVM and HMM models.

Figure 5 shows the results of Experiment 3 where we analyze the relative importance of LDCRF for contextual prediction and multimodal integration by running a new set of experiments where only the multimodal integrator changes. The ROC curves in this figure show that LDCRF model outperforms all three other models. This demonstrates the superiority of LDCRF for the multimodal integration task. Also, by comparing Figures 4 and 5, we can see that both SVM and HMM curves improve, confirming the utility of LDCRF as a contextual predictor.

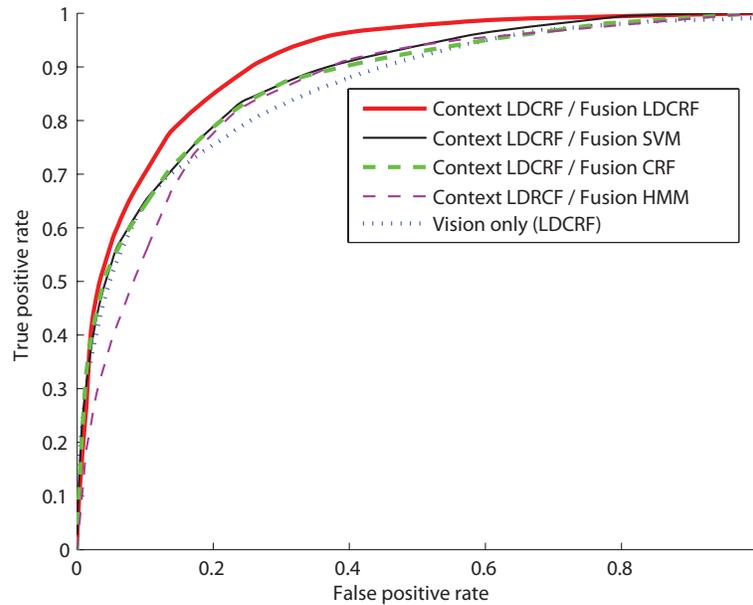


Fig. 5. Results from Experiment 3 analyzing the relative importance of LDCRF for contextual prediction and multimodal integration. Note that the contextual predictor is the same for all four cases and only the multimodal integrator changes in each case. This result demonstrates the superiority of LDCRF for the multimodal integration task and by comparing with Figures 4, we can see that both SVM and HMM curves improve, confirming the utility of LDCRF as a contextual predictor.

6 Conclusion

In this paper, we investigated how dialog context from an embodied conversational agent (ECA) can improve visual recognition of eye gestures. We proposed a new framework for contextual recognition based on Latent-Dynamic Conditional Random Field (LDCRF) models to learn the sub-structure and external dynamic of contextual cues. Our experiments showed that adding contextual information improves visual recognition of eye gestures and demonstrated that LDCRF models for context-based recognition outperform Support Vector Machines, Hidden Markov Models, and Conditional Random Fields for our visual feedback recognition tasks.

References

1. Morency, L.P., Christoudias, C.M., Darrell, T.: Recognizing gaze aversion gestures in embodied conversational discourse. In: Proceedings of the International Conference on Multi-modal Interfaces, Banff, Canada (2006)
2. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. Technical Report MIT-CSAIL-TR-2007-002, MIT CSAIL (2007)

3. Kendon, A.: Some functions of gaze direction in social interaction. *Acta Psychologica* **26** (1967) 22–63
4. Traum, D., Rickel, J.: Embodied agents for multi-party dialogue in immersive virtual worlds. In: *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*. (2002) 766–773
5. Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In: *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*. (2001) 301–308
6. Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., Hagita, N.: Messages embedded in gaze of interface agents — impression management with agent's gaze. In: *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*. (2002) 41–48
7. Velichkovsky, B.M., Hansen, J.P.: New technological windows in mind: There is more in eyes and brains for human-computer interaction. In: *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*. (1996)
8. Qvarfordt, P., Zhai, S.: Conversing with the user based on eye-gaze patterns. In: *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*. (2005) 221–230
9. Li, M., Selker, T.: Eye pattern analysis in intelligent virtual agents. In: *Conference on Intelligent Virtual Agents (IVA02)*. (2001) 23–35
10. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: *IEEE Intl. Conference on Computer Vision (ICCV)*, Nice, France (2003)
11. Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., Kobayashi, T.: A conversation robot using head gesture recognition as para-linguistic information. In: *Proceedings of 13th IEEE International Workshop on Robot and Human Communication, RO-MAN 2004*. (2004) 159–164
12. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labelling sequence data. In: *ICML*. (2001)
13. Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In: *NIPS*. (2004)
14. Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C.: Hidden conditional random fields for phone classification. In: *INTERSPEECH*. (2005)
15. Wang, S., Quattoni, A., Morency, L., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: *CVPR*. (2006)
16. Nakano, Reinstein, Stocky, Cassell, J.: Towards a model of face-to-face grounding. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan (2003)
17. Rich, Sidner, Lesh, N.: Collagen: Applying collaborative discourse theory to human-computer interaction. *AI Magazine, Special Issue on Intelligent User Interfaces* **22**(4) (2001) 15–25
18. Morency, L.P., Sidner, C., Lee, C., Darrell, T.: Contextual recognition of head gestures. In: *Proceedings of the International Conference on Multi-modal Interfaces*. (2005)
19. Kumar, S., Herbert., M.: Discriminative random fields: A framework for contextual interaction in classification. In: *ICCV*. (2003)
20. Vapnik, V.: *The nature of statistical learning theory*. Springer (1995)