Visual recognition of gestures in a meeting to detect when documents being talked about are missing

Hugo Tovar Lopez[1] and John Dowell[1]

[1] University College London, London, WC1. UK

j.dowell@cs.ucl.ac.uk

ABSTRACT. Meetings frequently involve discussion of documents and can be significantly affected if a document is absent. An agent system capable of spontaneously retrieving a document at the point it is needed would have to judge whether a meeting is talking about a particular document and whether that document is already present. We report the exploratory application of agent techniques for making these two judgements. To obtain examples from which an agent system can learn, we first conducted a study of participants making these judgements with video recordings of meetings. We then show that interactions between hands and paper documents in meetings can be used to recognise when a document being talked about is not to hand. The work demonstrates the potential for multimodal agent systems using these techniques to learn to perform specific, discourse-level tasks during meetings.

Keywords: Smart meeting room, multimodal agent, meetings video, object recognition, hand tracking, gesture recognition.

1 INTRODUCTION

Having the relevant documents to hand is key to whether meetings are properly informed and even whether the right decisions are made. Yet it is a common experience of attending meetings that a particular document being discussed or cited is not always present, perhaps because it could not have been anticipated that the discussion would take a certain turn or that a document would become so important to an agenda item. The same is true regardless of whether the meeting is co-located or distributed and whether documents are digital or paper. This absence of a needed document is one of the factors that can conspire against the success of meetings [12]; consider, for example, when a person at a meeting makes a critical or controversial claim citing their recall of a particular document.

Meetings and the activities they support have a domain of known documents such as minutes of previous meetings and documents such as reports, contracts, proposals etc. We contrast these kinds of document with unknown but relevant documents that might be gathered opportunistically, for example during an ideas generating activity [19]. We are focusing on the known and particular documents that a meeting might discuss.

Having the right documents to hand in a meeting was traditionally a task for administrator assistants, both by anticipating what documents would be needed and by following the discussion in a meeting and proactively retrieving a relevant document. Nowadays we seek to facilitate meetings and mitigate counter-acting factors with tools, but with the demand that their benefits must outweigh any downsides such as distraction effects [12]. Moreover the concept of the smart meeting room has been advanced as a space possessing ambient intelligence capable of functions like automatic summarization and detection of decision and action items [4]. We can envisage the future smart meeting room having an agent system that spontaneously retrieves a document being discussed which is absent.

An agent system with this capability would need to be able to judge whether a meeting is Talking_About_a_Document (the ‘TAD’ judgement) and whether that Document_is_Not_There (the ‘DNT’ judgement). It would retrieve the absent document or the most likely documents from a digital repository, using search terms from the dialogue proximal to where discussion of a document was detected. This agent system is most similar to the implicit querying, just-in-time information retrieval system for meetings investigated by Popescu-Belis and colleagues [24]. Their proposed system monitors conversations for pre-defined keywords informed by topic-modelling and uses these terms to search local repositories and the web for the most relevant documents (including unknown documents) and presents a continuously updating list of search results to the meeting participants. By contrast, the system we envisage uses multiple modalities to monitor for discussion of known documents and only offers documents that the meeting doesn’t already have.

To assess how amenable are the TAD and DNT judgement tasks for an agent system to perform, we can first ask how well do people do them? Since both involve cognitive and socio-cultural aspects, we can expect people to not always agree. As well as providing a performance baseline, such an enquiry would be able to identify the sources of information that a person uses and which an agent system might also exploit. We first report a study in which participants watched videos of meetings and performed both the TAD and DNT judgement tasks. The videos were selected from the AMI corpus of meetings videos [21]. Participants' judgements were used as the ground truth annotations of the videos to train agent techniques to make the same judgements.

An agent system making DNT judgements should base its judgements primarily on the visible activity and behaviours of participants in the meeting. Visual data can also be used for making TAD judgements. We report our exploratory application of computer vision techniques that learn to make DNT and TAD judgements. Visual object recognition and tracking are used to train neural network based classifiers. With selected videos from the AMI corpus, the classifiers are trained first to recognize and track hands and paper documents, second to recognize types of interaction between hands and paper, and third to make the TAD and DNT judgements. Other applications of computer vision techniques to meetings have invariably concerned analysis of visual focus of attention of participants [3] or social signals [30].

2 People performing the TAD and DNT judgement tasks

To understand how and how well people perform the TAD ('Talking-About-a-Document') and DNT ('Document-Not-There') judgements, we performed a study with ten participants watching videos of meetings. We wanted to establish how consistently would different people make the same judgements, what were the high level cues they were using, and the predicted value for the meeting of being given that document. Further, we wanted a dataset of meetings videos annotated with 'expert' judgements that would provide a ground truth for exploring the application of agent system techniques.

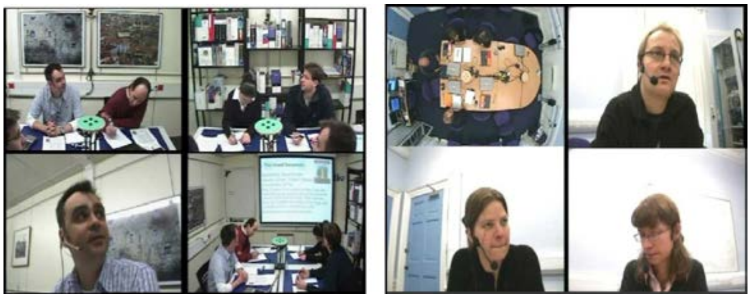


Fig. 1. Example views from multiple cameras for two AMI meetings.

Ten AMI videos were selected, each approximately half a minute long and with varying views from multiple cameras (Fig. 1), together representing a variety of meeting situations such as presentations and discussions. Each participant is recorded by two microphones and also available is a transcript of the meeting that is punctuated but not grammatically improved; it also includes non-word vocal sounds. There were explicit and implicit references to documents, discourse without reference to documents, explicit and implied statements about the presence of a document, and visual evidence for the presence and absence of documents (i.e. printed documents, projection screen, laptops).

Multi-choice questions about the TAD and DNT judgements were created through preliminary trials with 'observer' participants. Ten of these participants watched each of the videos with accompanying transcript and answered the open ended questions (a) 'Do you think this specific fragment is referring to a document*? (*As a document, consider any physical or electronic media containing text such as paper sheet, book, file, email or website)', and (b) 'If there is reference to a document, is this document currently present?' Participants were recruited through crowdsourcing and carried out the tasks online. Their responses were merged to produce a preliminary list of multi-choice answers. For example, responses to question (a) included: '(i) Yes: It has been explicitly mentioned. (ii) Yes: They are talking about information that seems to be present in a document', etc. Responses to question (b) included: '(i) Yes: The document has been explicitly mentioned as currently present; (ii) Yes: The document is displayed on a personal screen, (iii) No: The speaker refers to a document in the past tense', etc. A new group of ten observer participants watched the videos and answered the multi-choice questions with the option of providing additional answers to the questions. Very few additional answer categories were submitted. A final panel of ten 'test' participants were recruited again by crowdsourcing, all having experience of attending meetings. They were shown the videos with transcripts and asked to answer each of the questions with the aggregated multi-choice answers.

The group of ten test participants were 74.4% consistent in their TAD judgements concerning whether a document was mentioned (calculated by dividing the number of agreements by the number of judgements). In the DNT judgements concerning whether the document was not there in the meeting, their agreement was 74.2%. However, as the percentage of overall agreement (POA) doesn't take chance into account, we are interested in calculating the kappa coefficient [10] as a chanceadjusted measure of agreement, which can take a value in the range of -1 (i.e. perfect disagreement) to 1 (i.e. perfect agreement), where zero is the chance probability. In the situation of having multiple raters not restricted to assign a certain number of cases to each category, free-marginal multi-rater kappa (κ_{free}) [26] is recommended [7]. For TAD judgements, the inter-rater reliability κ_{free} is 0.488, while for DNT judgements κ_{free} is 0.484. A positive κ_{free} coefficient means there is a level of agreement above chance and in this case it equates to a moderate agreement level [18]. Even though the overall agreement on each judgement was substantial, the judgements are independent, in other words, a positive judgement in the TAD task does not make a positive judgement in the DNT task more likely.

The high-level cues used by participants to make their judgements were elicited from their answers to the multichoice questions. Participants were found to rely on the dialogue in 78% of TAD judgements about whether the meeting was talking about a document whereas DNT judgements about whether the document was absent relied evenly on both auditory and visual cues.

The study confirms that observers of a meeting can judge whether a document is being discussed and whether it is present. It confirms that they are able to do this with an acceptable level of agreement which therefore warrants using their video annotations as ground truth for training an agent system to perform this task. The study also provides insight into the cues people use to make the TAD and DNT judgements. It confirms that people use both visual and verbal evidence for performing the tasks and combine the judgements they make with each. Our exploration of the use of multimodal agent techniques to performing the tasks use the same cues, although the transcribed speech is only used for the TAD task. We

3 Computer vision method for the TAD and DNT judgement tasks

Our participants used the views of meetings in the videos, and in particular the views of documents being manipulated by hands, to judge whether the meeting was talking about a document (the TAD task) and whether that document was not there in the meeting (the DNT task).

We now describe the exploratory development of trained computer vision methods for performing these same judgement tasks. ‘Automatic’ Human-Object Interaction analysis has not been pursued previously in relation to meetings analysis, other than recognising if participants are looking at a whiteboard or a shared screen as part of dealing with the ‘visual focus of attention’ problem. Generally, Human-Object Interaction recognition relies on the static pose of both human and object to infer if there is an interaction between them. Gupta et al extended this approach to analyses involving motion [16]. For example, they trained a model by supervised learning to recognize from videos whether people were drinking or making a phone call in interaction with different objects including a cup and a phone. The almost perfect results they achieved were unsurprising given the diversity of objects and interactions they studied.

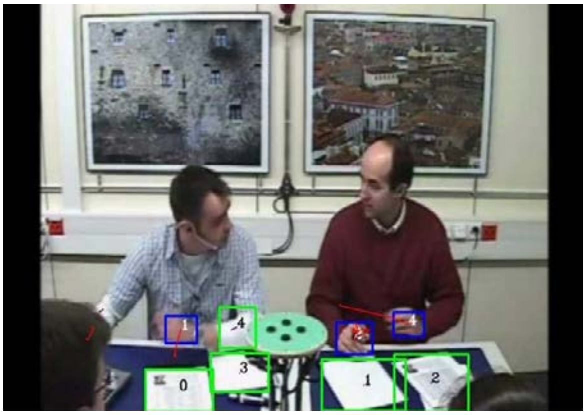
The meetings agent detects participants’ hands and paper documents, tracking their movements, then recognizing those movements as particular kinds of interaction between hands and paper. The association of those interactions with the two judgement tasks, TAD and DNT, can then be learnt from the ground truth annotations obtained from our observer participants.

3.1 Object detection and tracking

Detection of hands and paper documents in the videos has been implemented with C++ and OpenCV, an open source computer vision library that includes standard image processing techniques and analysis methods. Once hands and paper have been detected for each frame, they are tracked across frames. The tracking copes both with new hands and papers appearing and with existing ones disappearing due to occlusion.

To assess the effectiveness of the tracker, an epoch (16s long) from the IB4010 meeting video was used. Annotation of the ground truth location for each object was carried out by a volunteer using the Viper video annotation tool [11]. Figure 2 shows a typical frame containing several labelled hands and paper documents, including one hand that is currently occluded. The trajectory of each hand as determined by the tracker is shown. The performance of the tracker in locating hand #1 is shown in Figure 3 in terms of the relative displacement of the hand (measured in pixels) in each successive frame. The ground truth annotated data are also shown for comparison. Although the automatic tracked data has already been smoothed to remove false peaks, it still remains highly noisy. This error should be significantly reduced by implementing more advanced tracking techniques [32].

Assessment of the tracker’s performance in terms of the CLEAR MOT (Multiple Object Tracking) Metrics [6] is given in Table 1. The MOTP value for the hands and paper tracker is very high. The miss rate refers to objects not detected and is acceptable for this tracker, particularly so for paper. The false positive rate (FPR) expresses how many of the detected objects were incorrect and is high for the hand+paper tracker, a consequence of visual elements sharing the same pixel colours as hands and paper.



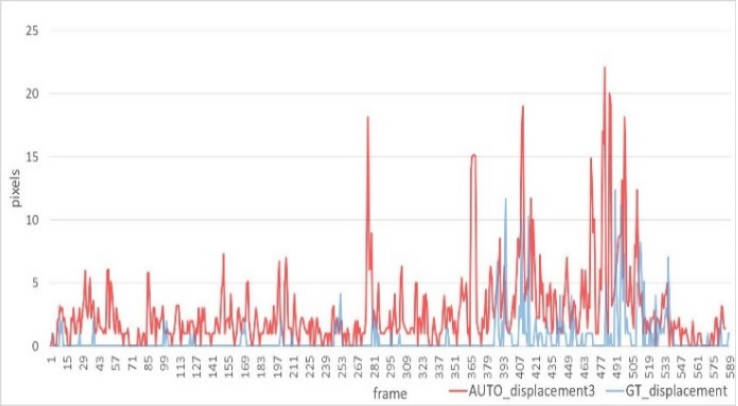


Fig. 2. Hands and paper tracking. Red lines show the trajectory of each hand. The red line on the lower left denotes a hand that is currently occluded but that has been tracked in previous frames. **Fig. 3.** Tracking of hand #1 over selected meeting epoch. Displacement of hand in each successive frame (pixels) from automatic tracking (AUTO) and manual annotation (Ground Truth GT).

Table 1. Evaluation of the tracker for hands and paper: MOTP (Multiple Object Tracking Precision); Miss rate (objects not detected). FPR (False Positive Rate); Mismatches (object confusion during tracking); MOTA (Multiple Object Tracking Accuracy).

Object	**MOTP** (pixels)	**Miss rate** (%)	**FPR** (%)	**Mismatches**	**MOTA** (%)
Hand	9.5	10.4	31.8	0	57.8
Paper	5.6	12.1	32.1	0	55.75

This rate can therefore be improved if shape detection techniques [23] and template matching are used[17]. Mismatches count the ratio of objects incorrectly matched during tracking. The perfect mismatch values for the tracker confirm the assumption that tracked hands and documents will be in a similar position in consecutive frames. The multiple object tracking accuracy (MOTA) metric is an aggregation over the miss rate, FPR and mismatches and conveys the tracker's overall strengths. The indifferent MOTA scores are due to the high FPRs and the impact of occlusions on the miss rate.

3.2 Hand+Paper interaction meta-features

Being able to track hands and papers in a meeting is a preliminary to automatically recognizing types of interactions between hands and paper. We assume there are few different types of such interactions, that they have a set of distinctive meta-features, and that an agent system can be trained to recognize them. We further assume that those interaction types associate systematically with the two judgements (TAD and DNT) and that an agent system can learn to make those judgements using recognized hand-paper interactions.

Three short videos (each of approximately 20s) were selected from the IB4010 meeting of the AMI Corpus (hereafter, Episodes A, B and C) and visually inspected for hand+paper interactions. Three interaction types were apparent in these episodes:

- GRABBING_STATIC: a static hand holding paper;
- GRABBING_MOVING: a moving hand holding paper;
- POINTING_READING: a finger points and touches a paper during reading.

Three meta-features were hypothesized as discriminating between the three interaction types

- Intersected area between hand and paper: the percentage of overlap between both objects, 0% being no intersection and 100% complete intersection.
- Inter-distance between hand and paper, the distance between both objects' bounding box centers, given in pixels.
- Self-displacement of hand and paper: the number of pixels the center of an object's bounding box moved from the previous to the current frame.

To investigate the regularity of these meta-features with the interaction types, the three video episodes were first annotated for: (i) the location of one relevant hand and one paper; (ii) the type of interaction between the hand and paper from the three categories; (iii) epochs where participants are talking about a document; (iv) epochs where the discussed document is available. Annotation of (i) and (ii) was carried out by a single volunteer whilst annotations of (iii) and (iv) were carried out by 10 crowdsourced volunteers because of the greater degree of individual judgement involved.

A hand+paper dyad from each video episode is analysed in relation to the interaction types and meta-features. Figure 4 shows (a) the area of overlap, and (b) the inter-distance for one hand+paper dyad in Episode A. Both ground truth annotated inter-distance and the inter-distances generated automatically by the hands+papers tracker are shown. Four instances of interactions of two different types (POINTING_READING occurring three times, GRABBING_MOVING occurring once) are mapped in the figure. Similar charts were generated for the other two meta-features of intersected area and self-displacement.

3.3 Developing a Hand+Paper interactions classifier

The question raised by the assaying of interaction meta-features from the detection and tracking of hands and papers, is whether there is a systematic relationship of those spatial data with the different interaction types that is amenable to machine learning. For example, in Figure 4 is it possible that a consistent pattern in inter-distance variation is associated characteristically with POINTING_READING which a classifier could be trained to identify?

We trained a neural network to classify every instance of interaction from the three videos of meetings episodes over every combination of hand and paper dyads from the set of hands and the set of papers. The neural network was trained with the annotated data and tested separately with the annotated spatial data and the data generated by our hands+papers tracker. We selected one hand+paper dyad from Episode A and an epoch consisting of 587 frames. Each frame was manually annotated for four meta-features (i.e. hand displacement h_d, paper displacement p_d, intersected area between hand and paper $(h, p)_a$, and the inter-distance $(h, p)_d$ between them). Hand+paper interactions were manually annotated; GRABBING_MOVING and POINTING_READING interactions but not GRABBING_STATIC interactions were identified in the epoch.

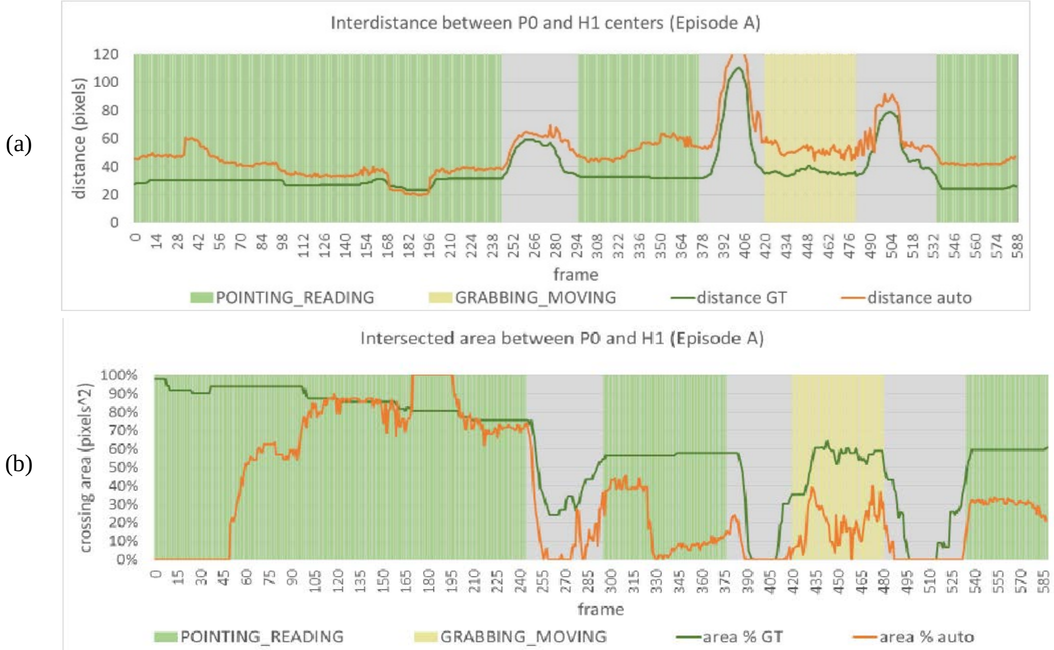


Fig. 4. Plots of (a) intersected area and (b) inter-distance between a Hand and Paper dyad on a succession of frames for both manually annotated (GT) and tracker generated (auto) data. Annotated Hand+Paper interactions (POINTING-READING; GRABBING_MOVING) are overlaid.

The data were randomly grouped as 70% for training, 15% for validation and 15% for testing. Table 2 shows the evaluation results from testing with both annotated and automatically tracked data. For the annotated data, the accuracy of classification is 97.7% for both types of interaction occurring in the selected epoch. The sensitivity rates are very high and the False Positive Rates extremely low. Accuracy deteriorates markedly when the automatically generated tracking data is used. This is consistent with the mixed precision (MOTP) and accuracy (MOTA) evaluation results obtained for the tracker, confirming the need to introduce more sophisticated image analysis techniques into the tracker to more accurately detect hands and paper. Nevertheless the potential for the classifier to recognise hand+paper interactions is clear from the performance with the annotated hand and paper location data.

Table 2. Evaluation results for the hand+paper (h,p) interactions classifiers. Results for manually annotated data and data produced by the trackers are shown. Inputs are hand and paper (h, p) dyad spatial features.

Classifier	**Input** (h,p)	**Accuracy %**	**Sensitivity %**	**FPR %**
GRABBING_ MOVING	annotation	97.7	97.6	0.0
POINTING_READING	annotation	97.7	100.0	3.3
(GRABBING_ STATIC)	annotation	n/a	n/a	n/a
GRABBING_ MOVING	tracker	60.1	95.3	82.6
POINTING_READING	tracker	55.2	44.2	1.7
(GRABBING_STATIC)	tracker	n/a	n/a	n/a

3.4 Developing two classifiers for the TAD and DNT judgements

The ability to recognise hand+paper interactions in a meeting implies that these interactions can be used for the two judgement tasks, TAD and DNT that together represent the decision of whether a meeting is talking about a document that is not to hand. Two neural network classifiers were trained, each performing one of these tasks. The classifiers were trained with the same 587 frame epoch as the interactions classifier and trained with the same selection of 70% of the data. Each classifier is capable of single-layered and of hierarchical recognition approaches [2], see Table 3.

- Single-layer approach: Uses hand and paper dyad (h, p) spatial features as input: hand displacement hd, paper displacement pd, intersected area between hand and paper (h, p)a, and the inter-distance (h, p)d between them. Outputs a binary judgement (high-level) for TAD and another for DNT.
- Hierarchical approach: hand+paper interactions from the interactions classifier and uses these to output a binary judgement (high level) for the TAD and DNT tasks.

The evaluation results for the two classifiers are shown in Table 4. For the TAD and DNT judgements with the annotated hand and paper spatial data, the single-layer approach obtains a higher accuracy (94.3% for TAD and 93.2% for DNT) than the hierarchical (70.5% and 81.8%), although for DNT this includes a false positive rate of 20%, in contrast with the 0% for the hierarchical. When the automatically tracked hand and paper data rather than the annotated hand and paper spatial data are used with the single layer classifier, the accuracy again deteriorates to just above chance as a consequence of the tracker's poor accuracy.

Table 3. Abstraction layers for recognition and classification of hand-paper interactions for the TAD and DNT judgement tasks.

Layer	**Analysis**	**Judgement**	**Features to extract**
High level	TAD & DNT	Document talked about? Document not there?	output
Mid-level	Hand and Paper interactions	What type of interactions are there between hands and papers?	Intersected area, inter- distance, and self-displacement
Low level	Hand and paper detection and tracking	Are hands and papers visible and where are they moving?	Bounding box positions and areas
Input	n/a	n/a	Video frames (pixels)

Table 4. TAD classifier and DNT classifier evaluation results with manually annotated data and automatically tracked data.

Classifier	**Input**	**Accuracy %**	**Sensitivity %**	**FPR %**
TAD (single-layer)	(h,p) spatial annotation	94.3	93.9	5.5
DNT (single-layer)	(h,p) spatial annotation	93.2	97.1	20.0
TAD (hierarchical)	h+p interactions	70.5	66.7	27.9
DNT (hierarchical)	h+p interactions	81.8	81.8	0.0
TAD (single-layer)	(h,p) spatial tracker	62.2	50.1	2.7
DNT (single-layer)	(h,p) spatial tracker	68.8	76.3	68.4

4 Discussion

People observing meetings can be reasonably consistent in their judgements about whether a document is being discussed and whether that document is present. This was the finding from our study with observer participants who made the TAD (is the meeting Talking-About-a-Document?) and DNT (is the Document-Not-There?) judgements. They used both the spoken discourse and the visible behaviour of people manipulating paper documents in the meetings to make these judgements.

We have also shown that there are distinct types of spatial interaction between hands and paper documents in meeting situations. The spatial meta-features we used to distinguish these interactions have been confirmed as good descriptors for classifying interactions. The interactions we identified are POINTING_READING, GRABBING_ MOVING, GRABBING STATIC and also NO-INTERACTION. The classifier we trained with annotated AMI meetings recordings has shown that it is possible to classify hand+paper interactions with an accuracy of 97.7% when using the manually annotated object tracking.

The results demonstrate that visually tracked hands and papers in a meeting can be used for the TAD and DNT judgements. A discrete classifier was trained for each judgement. Each was both single layer and hierarchical, capable of making the judgements directly from training with the hand and paper spatial data, and hierarchically with input of recognised hand+paper interactions. Evaluation found that the single layer accuracy exceeds the hierarchical (94.3% and 93.2% for TAD and DNT respectively, against 70.5% and 81.8% for the hierarchical model). These results were obtained with a subset of the AMI corpus and need to be replicated by re-training with a larger data set.

The classifiers' accuracies fell markedly with the automatically tracked data. The limitations of the tracker were evident in its inconsistent MOTP and MOTA evaluation scores, particularly its false positives rate. More sophisticated tracking algorithms [25] can be directly substituted into the tracker to replace the simple gausssian distribution mechanism. Optimising the tracker was a subsidiary aim in this work, our priority was to show that tracking of hands and paper documents is possible and that it can be exploited by the higher level judgement tasks. This is evidenced in the comparison of TAD and DNT judgements with the manually annotated and automatically tracked hands and paper.

The superiority of the single-layered approach is in part due to the interaction classifier discarding spatial feature information that is not relevant for its own purposes but may be useful for the subsequent TAD and DNT judgements. Since we can obtain high accuracy results for the TAD and DNT judgements using a single-layered approach (i.e. using the tracking information alone), it could be argued that there is no need to recognise hand and paper interactions. However, identifying the interactions enables explanation of the meeting participants' behaviour with documents. Moreover, the hand+paper interactions could be used as input to the text analysis modality as an intermediate analysis product to improve its modality specific judgements and to support fusion of modalities more extensively.

This study reveals how such a modality-specific technique can be exploited by an agent system [22]. The results indicate that the methods are usable and with development would be deployable for spontaneous retrieval of documents. An agent system using multiple modalities for the same judgements would achieve a higher performance by combining the judgements. That fusion process could occur earlier, with intermediate analysis products, or later with decision outcomes. The capability of the techniques revealed in this study would therefore improve not only with the modality specific improvements described, but also through their eventual fusion.

An agent system spontaneously retrieving an absent document for a meeting would need to decide whether and when that document should be presented. For example, alerting a meeting when a speaker has finished speaking, using a visual cue to indicate which document is being offered and perhaps which meeting topic this relates to, might make the distraction or interruption more acceptable [20]. This sort of judgement is often performed exquisitely well by people and it is feasible that an agent system would be capable of making this judgement too, given the capability we have found for judging whether a meeting is discussing an absent document.

5 REFERENCES

1. Aberdeen, J., Burger, J., Connolly, D., Roberts, S., and Vilain, M., 1993. Mitre-bedford: Description of the alembic system as used for muc-5. In *5th conference on Message understanding* Association for Computational Linguistics, 137-146.
2. Aggarwal, J.K. and Ryoo, M.S., 2011. Human activity analysis: A review. *ACM Computing Surveys (CSUR) 43*, 3, 16.
3. Ba, S.O. and Odobez, J.-M., 2011. Multiperson visual focus of attention from head pose and meeting contextual cues. *Ieee Transactions on Pattern Analysis and Machine Intelligence 33*, 1, 101-116.
4. Banerjee, S. and Rudnicky, A.I., 2007. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Intelligent User Interfaces IUI2007*, 151-159.
5. Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S.E., and Widom, J., 2009. Swoosh: A generic approach to entity resolution. *The VLDB Journal—The International Journal on Very Large Data Bases 18*, 1, 255-276.

6. Bernardin, K. and Stiefelhagen, R., 2008. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing 2008*, 1, 1-10.
7. Brennan, R.L. and Prediger, D.J., 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement 41*, 3, 687-699.
8. Chen, Z.Q., Kalashnikov, D.V., and Mehrotra, S., 2009. Exploiting context analysis for combining multiple entity resolution systems. *Acm Sigmod/Pods 2009 Conference*, 207-218.
9. Chinchor, N. and Robinson, P., 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, 29.
10. Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychosocial measurement, 20, 37-46.
11. Doermann, D. and Mihalcik, D., 2000. Tools and techniques for video performance evaluation. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 167-170.
12. Ehlen, P., Purver, M., Niekrasz, J., Lee, K., and Peters, S., 2008. Meeting adjourned: Off-line learning interfaces for automatic meeting understanding. In *Intelligent User Interfaces IUI2008* ACM, 276-284.
13. Galliano, S., Gravier, G., and Chaubard, L., 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, 2583-2586.
14. Garofolo, J.S., Laprun, C., Michel, M., Stanford, V.M., and Tabassi, E., 2004. The nist meeting room pilot corpus. In *LREC* Citeseer.
15. Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L., 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop* Association for Computational Linguistics, 92-100.
16. Gupta, A., Kembhavi, A., and Davis, L.S., 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *Ieee Transactions on Pattern Analysis and Machine Intelligence 31*, 10 (Oct), 1775-1789. DOI= http://dx.doi.org/10.1109/Tpami.2009.83.
17. Kulkarni, S., Manoj, H., David, S., Madumbu, V., and Kumar, Y.S., 2011. Robust hand gesture recognition system using motion templates. In *ITS Telecommunications (ITST), 2011 11th International Conference on* IEEE, 431-435.
18. Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *biometrics*, 159-174.
19. Li, N. and Dillenbourg, P., 2012. Designing conversation-context recommendation display to support opportunistic search in meetings. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia* ACM, 12.
20. Lopez-Tovar, H., Charalambous, A., and Dowell, J., 2015. Managing smartphone interruptions through adaptive modes and modulation of notifications. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* ACM, 296-299.
21. Mccowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., and Karaiskos, V., 2005. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*.
22. Mccowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D., 2005. Automatic analysis of multimodal group actions in meetings. *Ieee Transactions on Pattern Analysis and Machine Intelligence 27*, 3 (Mar), 305-317.
23. Ong, E.-J. and Bowden, R., 2004. A boosted classifier tree for hand shape detection. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on* IEEE, 889-894.
24. Popescu-Belis, A., Yazdani, M., Nanchen, A., and Garner, P.N., 2011. A just-in-time document retrieval system for dialogues or monologues. In *Proceedings of the SIGDIAL 2011 Conference* Association for Computational Linguistics, 350-352.
25. Pulford, G.W., 2005. Taxonomy of multiple target tracking methods. *Iee Proceedings-Radar Sonar and Navigation 152*, 5 (Oct), 291-304. DOI= http://dx.doi.org/10.1049/ip-rsn:20045064.
26. Randolph, J.J., 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.
27. Talburt, J.R., 2011. Entity resolution and information quality. *Entity Resolution and Information Quality*, 1-235.
28. Tur, G., Stolcke, A., Voss, L., Peters, S., Hakkani-Tur, D., Dowding, J., Favre, B., Fernandez, R., Frampton, M., Frandsen, M., Frederickson, C., Graciarena, M., Kintzing, D., Leveque, K., Mason, S., Niekrasz, J., Purver, M., Riedhammer, K., Shriberg, E., Tien, J., Vergyri, D., and Yang, F., 2010. The calo meeting assistant system. *Ieee Transactions on Audio Speech and Language Processing 18*, 6 (Aug), 1601-1611. DOI= http://dx.doi.org/10.1109/Tasl.2009.2038810.
29. Van Leeuwen, D.A. and Huijbregts, M., 2006. The ami speaker diarization system for nist rt06s meeting data. In *International Workshop on Machine Learning for Multimodal Interaction* Springer, 371-384.
30. Vinciarelli, A., Pantic, M., and Bourlard, H., 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing 27*, 12, 1743-1759.
31. Wilcock, G., 2009. Introduction to linguistic annotation and text analytics. *Synthesis Lectures on Human Language Technologies 2*, 1, 1-159
32. Yilmaz, A., Javed, O., and Shah, M., 2006. Object tracking: A survey. *ACM Computing Surveys 38*, 4.

adfa, p. 9, 2011.