Cognitive Multimodal Processing: from Signal to Behavior

Alexandros Potamianos
School of Electrical & Computer Engineering, National Tech. Univ. of Athens, Zografou 15780, Greece
potam@central.ntua.gr

ABSTRACT

Affective computing, social and behavioral signal processing are emerging research disciplines that attempt to automatically label the emotional, social and cognitive state of humans using features extracted from audio-visual streams. I argue that this monumental task cannot succeed unless the particularities of the human cognitive processing are incorporated into our models, especially given that often the quantities we are called to model are either biased cognitive abstractions of the real world or altogether fictional creations of our cognition. A variety of cognitive processes that make computational modeling especially challenging are outlined herein, notably: 1) (joint) attention and saliency, 2) common ground, conceptual semantic spaces and representation learning, 3) fusion across time, modalities and cognitive representation layers, and 4) dual-system processing (system one vs. system two) and cognitive decision nonlinearities. The grand challenges are outlined and examples are given illustrating how to design models that are both high-performing and respect basic cognitive organization principles. It is shown that such models can achieve good generalization and representation power, as well as model cognitive biases, a prerequisite for modeling and predicting human behavior.

Keywords

affective computing; social signal processing; behavioral signal processing; dual process theory; multimodal fusion; cognitive models; distributional semantic models; concept representations

1. INTRODUCTION

In the past decade, we have witnessed an explosion of research in the areas of multimodal signal processing, multimedia processing, and natural language processing, in an effort to bridge the gap between low-level feature representation and semantics, as well as, among semantics, behaviors and interaction. The early emphasis for these emerging research areas was on affective computing, especially trying to map low-level signal representation to discrete or continuous emotional space. The field has made tremendous progress in the past years, achieving good results in recognizing affect from audio, video, text and multimodal signals. Emotion recognition has found its way into the commercial world, being integrated into call center analytics solutions, text analytics and marketing (in the form of sentiment analysis). For a review of the field of affective computing see [7, 37].

More recently researchers turned their attention to higher level information, i.e., trying to map from low-level signal feature representations to attitudes, behaviors and interactional cues. The emerging field of social signal processing (SSP) automatically processes audio-visual streams in order to determine (in addition to emotion) personality, status, dominance, persuasion, rapport etc. Social signal processing concentrates on interaction and as such the basic maxims of human-human interaction are very relevant. Establishing common ground is an especially important maxim as highlighted by Vinciarelli et al. in [51]: "human-human communication is always socially situated and that discussions are not just facts but part of a larger social interplay" (more on this later). For a review of the field of SSP see [51, 36].

A related emerging field is that of behavioral signal processing (BSP) and the associated broader field of behavioral informatics. As defined in the SAIL lab at USC by Narayanan and colleagues: "Behavioral signal processing focuses on gathering, analyzing and modeling multimodal behavior signals, both overtly and covertly expressed ... in addition to processing objectively-specified behavioral content in richer ways (e.g., what someone said and did), behavioral signal processing entails automating a host of subjectively-specified entities such as those related to socio-emotional states of people (e.g., how negative or frustrated a person is; politeness; engagement etc)." Although the emphasis again here is on interaction, behavioral tracking can extend to all human activities. Note that a significant portion of the application of BSP are in the medical domain. For a review of behavioral informatics see [35].

A challenge for both the SSP and BSP fields are defining, labeling and annotating the high-level behaviors associated with human-human interaction. For this purpose experts in multimodal signal processing and machine learning work hand-in-hand with psychologists, clinicians and other do-

RFMIR '14, November 16 2014, Istanbul, Turkey
Copyright 2014 ACM 978-1-4503-0615-7/14/11...$15.00.
http://dx.doi.org/10.1145/2666253.2666264

main experts to transfer knowledge gained over years of labeling human behaviors to a machine readable code that is amenable to computational manipulations [4].

2. MAXIMS OF INTERACTION

Broadly based on the work of Tomasello [47, 48] (and others) human-human interaction can be represented as a three-step process: sharing attention, establishing common ground and forming shared goals (aka joint intentionality). Two prerequisites for successful human-human communication via joint intentionality are: 1) our ability to form a successful model of the cognitive state of people around us, i.e., decoding not only overt but also covert communication signals also referred to as "recursive mind-reading" and 2) establishing and building trust, a truly human trait. Affective computing, social and behavioral signal processing address the first prerequisite for successful communication; namely the ultimate aim of these disciplines is to build machines that can perform "mind-reading", i.e., understand the emotional, social and cognitive state of an individual, and going one step further identifying his/her intentions. SSP and BSP are also taking baby-steps towards identifying the correlates of trust that is essential for collaboration and cooperation, either directly or indirectly by measuring entrainment, engangement and joint intentions.

For many human-computer interaction scenarios, joint intentionality is not studied explicitly both because identifying intentions is a hard problem, but mostly because cooperation is often a given, i.e., computer and human are assumed to be working towards a common goal. Of course even for collaborative tasks the level of cooperation may vary, but this is beyond the scope of this paper. Next I focus on the first two steps of communication, namely: joint attention and common ground.

3. ATTENTION AND SALIENCY

Attentional mechanisms are widespread in biological organisms. Given the scarcity of cognitive processing resources and the limitation of human cognition (most) living organisms have developed (mainly low-level processing) algorithms to separate "foreground" from "background", i.e., signal from noise. Unlike computers, we are able to look without seeing; using a very focused attentional beam we may process only the small, most salient, parts of an image, a sound or a brochure, literally ignoring the rest. To summarize: "saliency detection is considered to be a key attentional mechanism that facilitates learning and survival by enabling organisms to focus their limited perceptual and cognitive resources on the most pertinent subset of the available sensory data" (cf. Wikipedia entry on saliency). Note that the attentional mechanism described above can be broadly defined as a *cognitive bias*, i.e., a limitation of human cognition that can lead to illogical deductions. For examples on how attention can be manipulated in stage magic see [27].

Saliency- and attention-based models have played a significant role in image and video processing in the past decade [26, 43, 5, 40, 11]. Using mostly low-level features such as color, intensity, motion, orientation etc., as well as some mid-level ones such as shapes, edges and contours, researchers were able to construct models that can accurately predict the focus of visual attention for images and videos, i.e., computer generated saliency maps. Saliency and attention is less pervasive in the audio/speech processing field; however, important finding are surfacing from neurocognition and cognitive science [33, 25]. High energy and spectral change are two important features for determining audio saliency. The notion that spectral change is salient for speech recognition applications has been captured explicitly via dynamic features (e.g., derivatives of MFCCs) as well as, via variable frame rate processing and frame dropping/attenuation [24, 10, 52]. Variable frame rate processing is motivated by the fact that regions of spectral change should be more finely sampled. Robust speech recognition focuses on the high-energy, most salient portions of speech for example: "missing feature theory" [34, 50], soft-feature decoding [38], attention-shift decoding [23, 21], island decoding [42]. In [20], the biological mechanisms of attention are investigated explicitly for speech applications.

Being able to model and predict what a human sees and hears in an audio-visual scene is the first step towards forming a cognitive model of that scene. Salient event detection using low-level cues from the audio, visual, and spoken language (transcription) modalities have proved very successful in identifying salient events in multimedia for a variety of applications. For example, attention-based algorithms can capture 60-80% of salient portions of an image, 75-80% of prominent syllables and words in read speech [20] and 70-90% of subjectively salient events in movies [11]. However, attention-based algorithms typically use only perceptually motivated low-level (frame-based) features and employ no high-level semantic information. Challenges still remain on: 1) extracting mid- and high-level feature extraction including incorporating semantics (scenes, objects, actions), 2) fusion of features over time and over modalities and 3) computational models for the fusion of the bottom-up (gestalt-based) and top-down (semantic-based) attentional mechanisms. Also applying these multimodal salient models to realistic human-human (especially) and human-computer interaction scenarios remains a challenge. Finally, identifying the dynamics of attention, i.e., constructing joint (interactional) attention models remains an open problem in this area.

4. COMMON GROUND AND CONCEPT REPRESENTATIONS

Establishing common ground is a prerequisite for successful communication. Negotiating interactional context and pragmatics is a complex task [9] that involves a wealth of information that has been encoded in our cognitive circuits. For humans, common ground often extends beyond the message itself: even the communication protocol proper (speech, gestures, body language) is a result of social convention and often has to be (re-)negotiated to establish common ground.

Humans form impressions, feelings, attitudes, beliefs and intentions under the constraints of the physical world that we live in (situational understanding) and also by mapping the physical/sensory world to an internal cognitive representation (embodied understanding). These constraints guide our cognitive models, specifically our semantic knowledge that forms the conceptual spaces in our brains. According to [14], concepts are cognitively organized in small dimensional subspaces that have good geometric properties and have a close correspondence to the physical world. For example, it is well established that the sense of touch is organized as a two-dimensional map of the body (down from three dimensions in the physical world). Similarly color is cognitively organized as a circle. The embodiment of the physical world in our cognition extends beyond concrete concepts to abstraction via analogy. Understanding can be modeled as activation of the appropriate semantic subspaces, as well as mapping between or constrains on these subspaces. Cognitive representations form the very basis of communication grounding so it is worth spending some time to analyze them in more depth.

This division of the conceptual space in numerous semantic "domains" might seem unnatural at first to a computer scientist accustomed to a single (typically metric) space that represents information. However, when the computational capabilities of the brain are closely examined, this extreme parallelization in processing is only natural. For example, in [13], it is vividly described how "the brain is a best-match computer; its massively parallel interconnected structure allows it to combine many factors in understanding a sentence or image" (cf. figure 1.1). Doing computations in parallel is also more efficient and achieves better performance given the processing speed of neurons. It is important to note that by breaking down computation in low dimensional subspaces, our cognition also battles the "curse of dimensionality," a problem that only the most advanced machine learning algorithms can successfully tackle. However, this massive parallelization also comes at a price. Elaborate fusion schemes have to be devised to combine intermediate results from all these subspaces, leading to higher interconnectivity, as well as cognitive biases (poor performance) when the fusion is suboptimal.

4.1 Why Cognitive Modeling?

Why would we want to create machines that actually mimic human cognition, including the errors (cognitive biases) that humans often make? For machine learning tasks that correspond to operations performed by low- and mid-level cognition, very high classification performance can be obtained by algorithms that are not motivated by cognitive representations. In fact, it is possible to build machines for specialized tasks that perform better than humans in recognizing patterns that can be objectively labeled by devising algorithms that deviate from cognitive processing, e.g., multi-talker speech recognition [22]. However, as we move to learning patterns that are the output of mid- and higher-level cognition, the labeling of such information is no longer unambiguous. If fact, we often do model the very errors, subjective beliefs and attitudes (cognitive biases) of a human, e.g., for emotion recognition. Thus, taking a human-centered approach becomes a necessity as we move from signals to behaviors and intentions. In essence, understanding behaviors and intentions requires the modeling of a network of interconnected "labels" that are produced at various layers of our cognitive systems, as well as modeling the mapping between these layers and labels. A pure machine learning approach will often fail simply because it is impossible to have access and objectively label this intermediate information, as well as associated modeling biases of these cognitive processes.

A second motivation for using cognitively-motivated computational models is performance. The two most successful machine learning (ML) paradigms in the multimedia signal, language and interaction processing areas are: 1) learning from task-specific annotated (or partially annotated) data using statistical models (e.g., gesture recognition, speech recognition) and 2) learning by imitation/example in a reinforcement learning paradigm (e.g., spoken dialogue systems, robot manipulation, goal planning). These approaches and associated models have significant merit and have achieved tremendous success in classification, recognition and tracking problems, sometimes even reaching superhuman levels of performance as discussed above. However, these models and paradigms have failed miserably in learning from very few examples, generalizing across tasks, and in forming plans and goals on the fly from extremely sparse observations, overall, their generalization power, learning rate and induction capabilities are poor. Why have we yet to build machines with these human-like capabilities? Partially because we have been asking the wrong question: trying to solve a (constrained) learning problem (i.e., mapping tokens to labels) rather than attacking the (general, unconstrained) problem (i.e., organizing all tokens and labels in a network of knowledge and then inferring token-label mapping on that network). We have also opted out of simple cognitively-motivated learning paradigms such as associative learning [44] and learning by analogy [32] for more complex machine learning algorithms that are mathematically more elegant and achieve good classification performance on constrained problems, but are less efficient for induction and learning from very few examples.

As discussed above, humans process and disambiguate multimodal cues, integrating low-, mid- and high-level cognitive functions, specifically using the (mostly) low-level machinery of cue selection via (joint) attention, the mid-level machinery of semantic disambiguation via common ground and shared conceptual representations, and the high-level machinery of intention reading. An essential part in this process is an extensive cognitive semantic/pragmatic representation, a network of concepts and their relations that form the very essence of common ground. Using these cognitive networks rapid learning from very few examples can be achieved.

The tremendous power of using a concept network for robust rapid learning is demonstrated in Fig. 1. The semantics of an unknown word are learned in a fully unsupervised manner from sentences using associative learning (word co-occurrence). Note that pure associative learning (blue line) requires hundreds of sentences to infer the meaning of an unknown word, while network-based associative learning (red line) only needs a few examples to get the gist of a word (0.5 correlation with human ratings averaged over 28 unknown words using only three examples). The power of the network comes from the integration of all knowledge (all words except the unknown ones), as well as the representation of information in conceptual subspaces with good geometric properties (approximately metric), where learning can be modeled as a contraction that converges to the true meaning of a word or concept. Network-based representations have similar motivation to neighborhood-based manifold learning algorithms, e.g., [41, 2] and representation learning [3]. For details on how these networks are formed and the semantic similarity metrics used, see [18]. For an extension to multimodal concept networks (containing words and images), see [17] based on the work of [6].

In situational human-human interaction, conceptual networks can be built and updated on the fly, creating com-

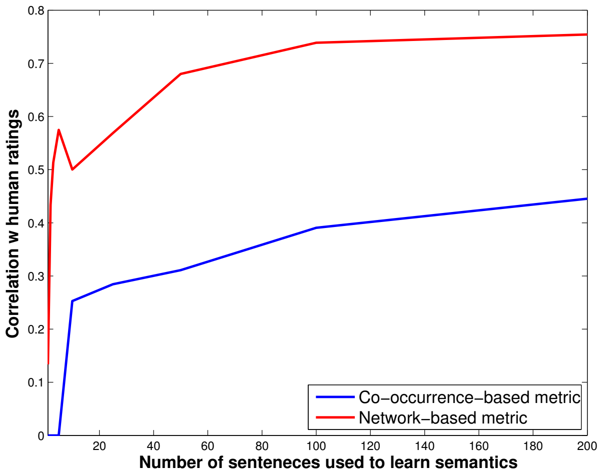


Figure 1: Learning the semantics of words with (red line) and without (blue line) a semantic representation in a fully unsupervised manner from a corpus. In the x-axis the number of sentences used for learning the semantics of a new word. Results are correlations with human judgments in a word pair semantic similarity task. One to three examples are enough to learn the semantics of a new word when using the semantic network proposed in [Iosif and Potamianos, 2012].

mon ground. They consist of associative networks of concepts (objects, words), actions, socio-affective states, and intentions that co-occur, e.g., "object" - "play" - "affective state: happy", as well as semantic relations between concepts and between actions determined using core semantic similarity algorithms and world knowledge. Understanding, learning and negotiating semantics and sharing intentions can be modeled on top of such a network. For example, the semantics of "object" will be learned via its image-based similarity with other concepts on the network, the fact that it is "play"-ed with etc. Computational models on how to map from an associative layer to semantic representations and from semantics to high-level labels such as affect is described in the next section. For more details on how to combine semantics and affect in cognitive and computational models of young children, see also the BabyAffect project `http://sites.google.com/site/babyaffectproject/`.

4.2 Global vs. Distributed Models

The question remains: can cognitively motivated machine learning algorithms perform well in classification and modeling tasks? In essence, does switching from serial to parallel processing and from a unified space to a union of fragmented low-dimensional subspaces compromise performance of our machine learning algorithms? In a recent publication, Athanasopoulou et al. [1] show that a semantic representation that uses a union of low-dimensional manifold distributional semantic models (DSMs) performs at least as well as the state-of-the-art. This is consistent with the intuition that as we move from low-level signal representations to higher-level semantic and behavioral representations, cognitively-motivated models should match or improve upon performance of traditional machine learning models. The manifold DSM construction process consists of four steps: 1) identify the domains that correspond to each of the low-dimensional manifolds (in this paper, a domain is created for each word or concept in the vocabulary and the domain is defined as the semantically similar words, i.e., the semantic neighborhood), 2) threshold the semantic similarity matrix to obtain a sparse representation, 3) run the dimensionality reduction algorithm for each domain to construct a very low-dimensional DSM for each domain, and 4) combine/fuse the local computation from each manifold DSMs to estimate measures of lexical relations. The process is depicted in Fig. 2 (from [1]), where v_i are the words for which each neighborhood/domains is built, P^{v_i} and X^{v_i} are the semantic similarity matrix (DSM) for each neighborhood before and after sparsity is applied, and $\tilde{S}^{v_i}$ is the low-dimensional manifold representation of each domain after multi-dimensional scaling is applied.

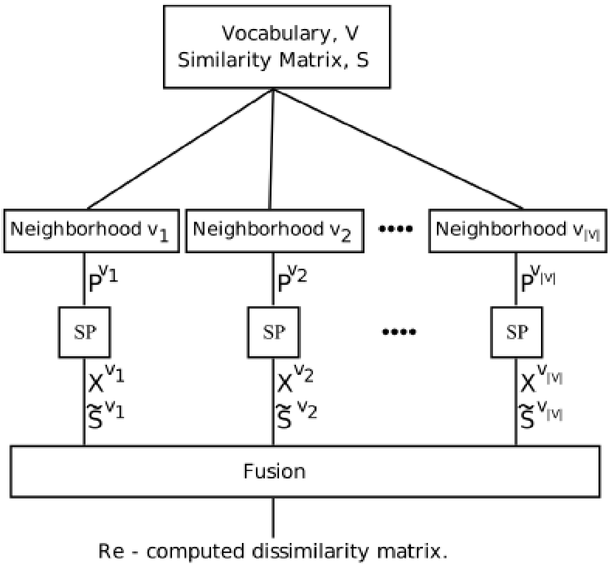


Figure 2: Low-dimensional manifold distributional semantic model. A domain is constructed from the semantic neighborhood of each word v_i, the semantic similarity matrix for each domain is thresholded to make it sparse (SP) and its dimensionality is reduced (typically down to 3 to 5). Semantic similarity is estimated as the fusion of estimates from local manifolds (from [Athanasopoulou et al., 2014]).

There are two important findings in this work: 1) even for very simplistic linear and non-linear (max) fusion of the local decisions in each subspace, this highly parallelized model performs at least as well as state-of-the art DSMs and word embedding models, and 2) best performance is achieved with subspaces of very low dimensionality of typically three dimensions (good performance is also achieved with only two dimensions); compare this with a dimension of a few hundred for typical DSM and word embedding models that operate on a single unified semantic space. For this task of matching semantic similarity ratings produced by humans (a task that included cognitive biases), following a cognitive approach both leads to high-performance and provides further validation for the assertion that conceptual spaces are fragmented and of very low dimension.

To conclude with the grand challenges for this section:

grounding, situational and embodied understanding all depend on a conceptual representation that has been acquired from our life experiences and negotiated over our social interactions. It makes little sense to approach social and behavioral signal processing problems "tabula rasa" as it is often attempted by the signal processing and machine learning community, i.e., here are some labels and features go ahead and construct a specific model for your task. Grounding exists only in the context of our semantic, affective and interactional cognitive representations and should be addressed as such. This poses the grand challenge of using "big data" to construct such cognitive representations, as well as defining the "topology" (unified vs. distributed) and processing logic (parallel/serial) of these representations. Taking a cognitively motivated approach is both a realistic and necessary step towards this goal. Cognitively-motivated conceptual representations can be used to design algorithms that achieve rapid learning and adaptation to new concepts and situations from very few examples (situational learning and understanding), provide grounding in interaction and problem-solving settings (negotiating common ground) and map to higher-level representation such as affect and other SSP/BSP correlates, as discussed next.

5. FROM SEMANTICS TO BEHAVIOR

How do we build machines that can map from low-level signal to high-level behavioral representations? Even if we were able to solve the multimodal understanding problem by mapping from signal(s) to semantics (a monumental task by itself), it would take us only half way there. Assuming that a conceptual representation is in place (see also discussion in the previous section), we proceed next to discuss how to model jointly semantics and affect.

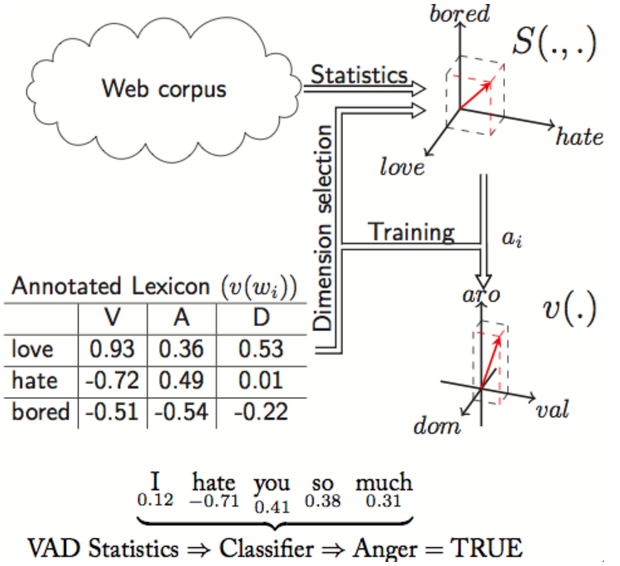
Given that the cognitive semantic space is both distributed and fragmented into subspaces the mapping from semantics to affective labels should also be distributed and fragmented. Semantic-affective models (SAM) [31, 49] are based on the assumption that semantic similarity implies affective similarity. Thus affective models can be simply constructed as mappings from semantic neighborhoods to affective scores. The basic idea behind these models is shown in Fig. 3.

In the SAM model proposed by Malandrakis et al. [31, 30], the valence (or any other affective label) of a word can be expressed as a linear combination of its semantic similarities to a set of seed words and their valence ratings:

$$\hat{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i \; v(w_i) \; d(w_i, w_j), \qquad (1)$$

where is w_j the unknown word, $w_1 ... w_N$ are the seed words $v(w_i)$ is the valence rating of word w_i, a_i is the (trainable) weight assigned to seed w_i, and $d(w_i, w_j)$ is the measure of semantic similarity between words w_i and w_j. Only the valence of (a few hundred) seed words needs to be labeled, the rest of the procedure (including training the semantic model) is fully unsupervised and language agnostic.

It is important to note that Eq. (1) is a general purpose solution on how to map between different layers of representations without assuming an underlying metric space (in our example the semantic space). In essence, the proposed model is a distributed mapping from semantic neighborhoods to affect (and potentially other labels). For example, the SAM model has been successfully applied to sentiment analysis ranking among the top systems in the Twitter data sentiment analysis task in the 2013 and 2014 SemEval campaigns [28], as well as for classifying discrete labels related to semantics and affect in [30] The model can be extended to also handle many-to-many mappings between multiple layers of cognitive representations.



	V	A	D
love	0.93	0.36	0.53
hate	-0.72	0.49	0.01
bored	-0.51	-0.54	-0.22

Figure 3: Semantic-affective model: 1) a semantic model is constructed in an unsupervised way from a web corpus, 2) the mapping between semantic and affective spaces is estimated using a few hundred labeled seed words (used for bootstrapping the model), 3) continuous affective scores for new words and n-grams are estimated, 4) scores are combined to form sentence level ratings (from [Malandrakis et al., 2014]).

The proposed model is a first step towards estimating affective and behavioral labels from a semantic representation. Although, good performance can be obtained for language and image processing applications, the challenge remains on how to apply this model to audio and video, where the segmentation of the stream into tokens is not straightforward. Also, the model works very well at estimating the affective content of single tokens (words, images); going from a single token to a sequence of tokens (e.g., word to sentences) is a hard open problem that is also discussed in Section 7.

6. DUAL SYSTEM PROCESSING

Dual process theory (sometimes referred to as system 1 vs. system 2 processing) states that there are two (somewhat separate) cognitive processes one that is rapid and provides impressions, feelings, and inclinations and one that is slower and provides beliefs, attitudes, and intentions (filtering system 1 input) [19, 45]. Early versions of this theory [12], separate associative vs. true reasoning, basically describing true reasoning as operations on top of a cognitive network of associations.

According to Kahneman, system 1 is always on, while system 2 turns on and off selectively based on the situation and the internal cognitive state. This highly nonlinear behavior significantly affects how beliefs, attitudes, and in-

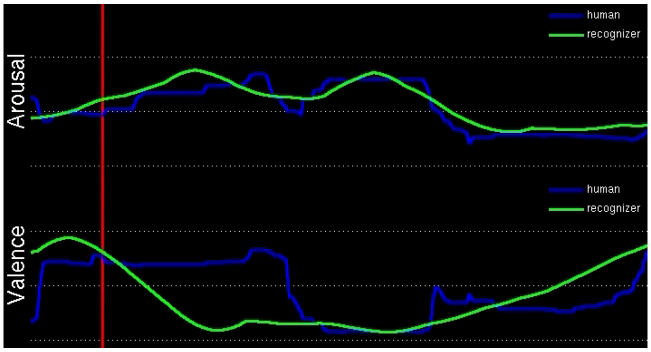


Figure 4: Continuous affective scores (arousal-top and valence-bottom) annotated by humans (blue line) and automatically estimated by the algorithm proposed in [Malandrakis et al., 2011] (green line) for a 2' clip of the movie "Finding Nemo". Note the abrupt drop of valence for the human rater vs. the gradual decrease for the machine.

tentions are formed in humans. It is thus impossible to create realistic models of complex human behavior without modeling the internal cognitive state. Psychologists, that have observed and codified human behavior in the lab, have come to this realization early on and have used states to codify attitudes/moods in their models. For example, the model for marriage described in [15] uses separate states for distancing, attraction, escalation, and problem acceptance. Dynamical systems are used to model the abrupt transition between those states.

An example on how linear-fusion, stateless computational models often fail to model abrupt transitions (in this case surprise) can be seen in Fig. 4 based on the work of Malandrakis et al. on continuous-time emotion tracking evoked from watching a movie [29]. Observe how for valence for the human annotators drops abruptly, while the hidden Markov models employed for emotion tracking in this work predict a very smooth transition (both in the first portion of the clip when valence decreases and towards the end where valence increases).

How can system 1 vs. system 2 processing be practically applied to signal and language processing problems? Inspiration can be drawn from the work of Iosif et al. on semantic networks [18]. In this work, estimation of semantic similarity between words is broken down as a two-step process: 1) first the semantic neighborhood for each word is selected, which is equivalent to the activation of a network of associations (priming) in human cognition (a system 1 process), and 2) semantic similarity is estimated via operations on the semantic neighborhoods (a system 2 process). The proposed algorithm despite its simplicity achieves state-of-the-art results for estimating word and image semantic similarity [17]. Designing models that are stateful and are able to predict cognitive biases, nonlinear logic, abrupt state transitions and surprise remains a major challenge for social and behavioral signal processing.

7. MULTIMODAL FUSION

Given the massively parallelized processing in our brains, fusion plays a central role in cognitive processing. Fusion is also an important and active research area in machine learning and signal processing, especially when merging features or representations from different modalities, e.g., audiovisual speech recognition [39], video summarization from aural, visual and language cues [11]. There are three basic type of fusion that are of interest: 1) multimodal fusion, i.e., fusion between modality-specific processing outputs and multimodal outputs, 2) fusion over time, i.e., how stimuli are integrated both within and across modalities, as well as how mid- and higher-level cognitive processes integrate information over time, e.g., going from frame and word level affective ratings to sentence level ones, and 3) fusion of top-down (data-driven) and bottom-up (semantic) processing, or in general fusion between different layers of cognitive and computational processing.

There has been significant research on fusion over time and across modalities for low- and mid-level cognitive processing and there are three generally agreed upon principles of multisensory integration [46]: 1) the spacial rule: stating that mutlisensory integration is more likely for co-located stimuli, 2) the temporal rule: stating that multisensory integration is more likely for stimuli that occur on or about the same time, and 3) the principle of inverse effectiveness: stating that multisensory integration is stronger when the corresponding unimodal stimuli are noisy or weak. Bayesian models have been applied successfully to the problem of multisensory integration tasks both by computer and cognitive scientists, e.g., [8]. When moving from low- to mid- and high-cognitive processes multimodal integration and integration over time often presents cognitive bias and nonlinearities. Thus, the problem of multimodal fusion for complex stimuli remains an open and actively researched problem. For example, the CogniMuse project investigates multimodal fusion and fusion over time for videos (especially movies and documentaries) using state-of-the-art multimedia and language processing technologies `http://cvsp.cs.ntua.gr/research/cognimuse/`.

Fusion of the output of various cognitive processes and especially fusion of top-down and bottom-up cognitive processes is a less researched and much harder problem. An example of such a fusion is how early bottom-up emotional responses often get suppressed in the amygdala by late top-down semantically driven input [16], e.g., a loud sound scares us, but fear gets suppressed when we figure out that the sound corresponds to a door slamming due to the wind. Another example of top-down and bottom-up fusion are the attentional mechanisms that are driven by low-level cues (bottom-up processing), but are corrected by top-down semantic information. In general, the selective activation of top-down processing based on our cognitive state (e.g., alertness) and the type of stimuli presented to us makes this type of fusion hard to analyze and model. Integrating semantic information and low-level features/cues in machine learning algorithms is also a challenge for computational models.

Designing efficient fusion algorithms is key for modeling the outputs of mid- and high-level cognitive processes, which is the goal for most social and behavioral signal processing tasks. Clearly top-down and bottom-up fusion, as well as dual-system cognitive processing introduces many nonlinearities that are not always easy to model using popular Bayesian statistical models. It thus remains a challenge to analyze cognitive fusion mechanisms and attempt to model them using novel machine learning models. This entails going beyond simple algorithms that employ (weighted) averages of outputs (across time, modalities and processes) and design algorithms that make often highly non-linear fusion decisions depending on our cognitive state, behaviors and intentions.

8. CONCLUSIONS

In this work, I have identified major challenges that lie ahead in the fields of affective, social and behavioral signal processing. I have argued that it is very improbable that one can successfully address these major challenges without taking into account the peculiarities of human cognition. The following challenges have been identified: 1) annotation of the mid- and high-level behaviors associated with human-human interaction, 2) attention and saliency modeling using mid- and high-level features (including semantics), as well as fusion model of top-down and bottom-up attentional mechanisms, 3) from signal to semantics: use "big data" to construct semantic cognitive representations that are distributed and low-dimensional, 4) from semantics to mid-/high-level cognitive SSP/BSP labels: estimate many-to-many mapping between semantics and other cognitive representation layers, 5) design models that are stateful and are able to predict cognitive biases, nonlinear logic, abrupt state transitions and surprise, and 6) design multimodal fusion algorithms that can exhibit nonlinear behavior and depend on cognitive state, behaviors and intentions. I have tried to provide examples on how to address each of the challenges according to my subjective view on these problems. However, all of these problems remain open and constitute fruitful research directions, in my view.

There are quite a few issues that have not been discussed here. The details on how to extend the proposed methods (applied here to words and images) to continuous time signals such as audio clips, speech and videos are omitted. Also, although the core challenges discussed here are directly applicable to human-human and human-machine interaction, models of interaction are not described explicitly. Finally, models of intentionality and joint intentionality are not reviewed. I hope to address these shortcomings in a future publication.

9. ACKNOWLEDGMENTS

This work was partially supported by the BabyAffect and CogniMuse projects under funds from the General Secretariat for Research and Technology (GSRT) of Greece, and the EU-IST FP7 SpeDial project. The author wishes to thank Dr. Elias Iosif for providing the results shown in Fig. 1 and for many useful discussions. Some of the ideas presented here have taken shaped through discussions with Prof. Shrikanth Narayanan at USC and at the Computational Audio Analysis Seminar at Dagstuhl in 2013. The author also wishes to thank Prof. Alessandro Vinciarelli and Prof. Dirk Heylen for their invitation to participate and submit a position paper to the social signal processing forward-looking workshop.

10. REFERENCES

[1] G. Athanasopoulou, E. Iosif, and A. Potamianos. Low-dimensional manifold distributional semantic models. In *Proc. COLING*, Dublin, Ireland, August 2014.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[4] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, 55(1):1–21, 2013.

[5] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet. A metric for no-reference video quality assessment for hd tv delivery based on saliency maps. In *Proc. ICME*, pages 1–5, 2011.

[6] E. Bruni, N.-K. Tran, and M. Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47, 2014.

[7] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[8] S. Deneve and A. Pouget. Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology-Paris*, 98(1):249–258, 2004.

[9] P. Dillenbourg and D. Traum. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151, 2006.

[10] S. Dimopoulos, A. Potamianos, E.-F. Lussier, and C.-H. Lee. Multiple time resolution analysis of speech signal using mce training with application to speech recognition. In *Proc. ICASSP*, pages 3801–3804, Taipei, Taiwan, 2009.

[11] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, Nov. 2013.

[12] J. S. B. Evans. Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4):451–468, 1984.

[13] J. Feldman. *From molecule to metaphor: A neural theory of language*. MIT press, 2008.

[14] P. Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004.

[15] J. Gottman, C. Swanson, and K. Swanson. A general systems theory of marriage: Nonlinear difference equation modeling of marital interaction. *Personality and social psychology review*, 6(4):326–340, 2002.

[16] D. B. Huron. *Sweet anticipation: Music and the psychology of expectation*. MIT press, 2006.

[17] E. Iosif. *Network-Based Distributional Semantic Models*. PhD thesis, Technical University of Crete, Chania, Greece, 2013.

[18] E. Iosif and A. Potamianos. Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering*, pages 1–31, 2013.

[19] D. Kahneman. A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697, 2003.

[20] O. Kalinli. *Biologically inspired auditory attention models with applications in speech and audio processing*. PhD thesis, UNIVERSITY OF SOUTHERN CALIFORNIA, 2009.

[21] O. Kalinli and S. S. Narayanan. Continuous speech recognition using attention shift decoding with soft decision. In *INTERSPEECH*, pages 1927–1930, 2009.

[22] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath. Super-human multi-talker speech recognition: the ibm 2006 speech separation challenge system. In *INTERSPEECH*, Pittsburgh, Pennsylvania, 2006.

[23] R. Kumaran, J. Bilmes, and K. Kirchhoff. Attention shift decoding for conversational speech recognition. In *INTERSPEECH*, pages 1493–1496, 2007.

[24] P. Le Cerf and D. Van Compernolle. A new variable frame analysis method for speech recognition. *IEEE Signal Processing Letters*, 1(12):185–187, 1994.

[25] A. K. Lee, E. Larson, R. K. Maddox, and B. G. Shinn-Cunningham. Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing research*, 2013.

[26] T. Liu, X. Feng, A. Reibman, and Y. Wang. Saliency inspired modeling of packet-loss visibility in decoded videos. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, pages 1–4, 2009.

[27] S. L. Macknik, M. King, J. Randi, A. Robbins, et al. Attention and awareness in stage magic: turning tricks into research. *Nature Reviews Neuroscience*, 9(11):871–879, 2008.

[28] N. Malandrakis, M. Falcone, C. Vaz, J. J. Bisogni, A. Potamianos, and S. Narayanan. Sail: Sentiment analysis using semantic similarity and contrast features. In *SemEval*, Dublin, Ireland, 2014.

[29] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi. A supervised approach to movie emotion tracking. In *Proc. ICASSP*, pages 2376–2379, Prague, 2011.

[30] N. Malandrakis, A. Potamianos, K. J. Hsu, K. N. Babeva, M. C. Feng, G. C. Davison, and S. Narayanan. Affective language model adaptation via corpus selection. In *Proc. ICASSP*, Florence, Italy, 2014.

[31] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392, Nov 2013.

[32] A. B. Markman and D. Gentner. The effects of alignability on memory. *Psychological Science*, pages 363–367, 1997.

[33] N. Mesgarani and E. F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012.

[34] A. C. Morris, M. P. Cooke, and P. D. Green. Some solution to the missing feature problem in data classification, with application to noise robust asr. In *Proc. ICASSP*, volume 2, pages 737–740, Las Vegas, Nevada, 1998.

[35] S. Narayanan and P. G. Georgiou. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, 101(5):1203, 2013.

[36] A. Pentland. Social signal processing. *IEEE Signal Processing Magazine*, 24(4):108, 2007.

[37] R. W. Picard. *Affective computing*. MIT press, 2000.

[38] A. Potamianos and V. Weerackody. Soft-feature decoding for speech recognition over wireless channels. In *Proc. ICASSP*, Salt Lake City, Utah, 2001.

[39] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[40] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1454–1461, 2009.

[41] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[42] T. N. Sainath. Island-driven search using broad phonetic classes. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 287–292, Merano, Italy, 2009.

[43] A. Shokoufandeh, I. Marsic, and S. J. Dickinson. View-based object recognition using saliency maps. *Image and Vision Computing*, 17(5):445–460, 1999.

[44] B. F. Skinner. *Verbal behavior*. BF Skinner Foundation, 2014.

[45] S. A. Sloman. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3, 1996.

[46] B. Stein and M. Meredith. *The merging of the sense*. Cambridge, MA: The MIT Press, 1993.

[47] M. Tomasello. *Origins of human communication*. MIT press Cambridge, 2008.

[48] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(05):675–691, 2005.

[49] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[50] M. Van Segbroeck and H. Van Hamme. Advances in missing feature techniques for robust large-vocabulary continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):123–137, 2011.

[51] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[52] D. Vlaj, B. Kotnik, Z. Kaciv, and B. Horvat. Usage of frame dropping and frame attenuation algorithms in automatic speech recognition systems. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8*, volume 2, pages 149–152 vol.2, Sept 2003.