Multimodal affect databases: collection, challenges, and chances

Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn. 2015. "Multimodal affect databases: collection, challenges, and chances." In The Oxford handbook of affective computing, edited by Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas, 323–33. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.005.

CHAPTER

23 Multimodal Affect Databases: Collection, Challenges, and Chances

Björn Schuller

Abstract

This chapter focuses on multimodal affect databases. After a short introduction, the collection of affective data is discussed in 10 steps highlighting methodological considerations and challenges of building new resources of multimodal data and affect labels. It then touches upon quality assessment of collected emotion corpora. A section is also dedicated to "saving labor" by sharing annotation between human and machine and reusing data. Then a selection of representative audiovisual and further multimodal databases is introduced. Finally, the chapter concludes with a discussion of controversial issues and future directions.

Key Words: multimodal affect databases, emotion corpora

Introduction

In order to train and test multimodal affect recognition and synthesis systems or analyze human affective behavior, data are needed. In fact, this is often considered to be one of the main bottlenecks, and a common opinion in machine learning is that there are "no data like more data." In this vein, this chapter aims to first, in Introduction (p. 323), give insight into the state of the art in multimodal affect data, then to outline requirements and discuss challenges. It then deals with steps toward preparation of such a multimodal database in Data Collection—Ten Steps Toward a Multimodal Affect Database (p. 324). Quality Assessment—Is This Really Joyful? (p. 327) deals with quality assessment of the annotation and weighting of raters. In Efficiency—How to Save Annotation Labor (p. 328) avenues toward reduction of human annotation efforts are shown, including active, semisupervised, and unsupervised learning. Selected examples of existing multimodal affect resources are shown in Existing Multimodal Resources—What's There? (p. 330), after which Conclusions and Future Avenues—Wrapping Up (p. 331) wraps up the discussion.

State of the Art

While there are increasingly multimodal recordings of affect displays available, to date these tend to be smaller in size and more often recorded in the lab than their unimodal counterparts (Gunes & Schuller, 2012). Until recently, the larger portion of affective databases comprised bimodal databases—usually of audiovisual nature. In addition, multicultural and multilingual data, the latter being especially important for text-based detection, are still considerably sparse. This is even truer when it comes to data in naturalistic or working system contexts. More "exotic" and richer combinations of data—such as physiological measures or speech alongside depth images—are, however, progressively more available. Further, more and more languages are covered and more natural databases with more complex affect labels are increasingly available to the community. In general, ever more

multimodal resources are to be expected (Schuller, Douglas-Cowie, Batliner, 2012).

Requirements

There are several prerequisites apart from the sheer quantity of the data required, and obtaining considerable amounts of data can be difficult and labo-intensive, since data must usually be labeled. The most relevant of these requirements include *quantity*; high diversity with respect to multiple factors such as the age, gender, and culture of subjects and the situational context; and reasonably balanced distribution of instances among classes or even along the range in case of continuous models. Next is the *quality* of the data, in the sense of adequate data, realistic and naturalistic data, and adhering to ideal capture conditions within acceptable parameters for levels of noise, reverberations, occlusions, and so on. A further requirement is *appropriate modeling* in terms of reasonable categorization or choice of appropriate dimensions together with well-defined mappings between potential models. The *labeling* demands *consistency* as well as additional provision of metainformation such as transcripts of spoken text and nonlinguistic vocalizations, context and events, individual labeler tracks, and so forth. There must be a high number of *annotators* (or *coders, labelers,* or *raters*)—ideally again with high demographic diversity, and the provision of a "gold standard" along with its reliability and potential additional emotion perception tests for verification of trustworthiness. Finally, the *release of data* requires additional considerations such as documentation of detailed recording conditions and metainformation on the subjects, provision of baseline recognition results by automatic engines for others to compare with, free release of the data with high accessibility, and suggested partitioning of the data into test, development, and training partitions to avoid arbitrary testing partitions.

Challenges

One of the major challenges in recording multimodal affective data is to obtain naturalistic displays of affect. The complex setups for multimodal recordings often require very careful control of lab conditions. However, when recording is done in the lab, one of the problems is the "observer's paradox": According to Labov (1984) and one's general intuition, the presence of the experimenter and the awareness of being recorded may influence the subject.

Another challenge typical for the recording of multimodal data is the synchronization of the multimodal capture streams, as these are often recorded by individual hardware devices and additionally may operate at different time scales. The sampling interval of these different hardware devices can partly be nonconstant, which makes the challenge of synchronization more demanding.

Then, a major challenge is the labeling of the data by a sufficient number of independent labelers or by the subjects themselves. In some multimodal setups, not all modalities' recorded data may be sufficiently informative for human labelers to make affect judgments. This may require self-assessment, ideally online, which again can be highly disruptive with respect to an awareness of being in an experimental setting. New methods for community or distributed annotation such as crowdsourcing (e. g., by Amazon Mechanical Turk) may be useful to obtain annotations with labels for a data instance with high variability of raters and to potentially reduce the cost of obtaining annotations (see Chapter 30, this volume, on crowdsourcing for affective computing).

Data Collection—Ten Steps Toward a Multimodal Affect Database

This section aims to highlight usual steps in preparing an affect database with a focus on multimodal data (a general introduction to the topic of affect databases is given in Chapter 26, this volume). The 10 aspects considered include ethics, the actual data acquisition by recording or "reusing" data, collection of metainformation, synchronization of the multimodal streams, choice of an appropriate affect modeling for the subsequent labeling while considering standards to be used, partitioning of the data for experimentation, human perception studies and baseline results (e.g., by automatic recognition), and the actual release to the community.

Considering Ethics

Ethical considerations represent one of the major issues with affect data collection. Affect can be very private, and subjects in the real world or in lab studies might not always agree with making genuine and spontaneous affect data available for study, in particular when it comes to video or audio recordings. With a gradual increase in the multimodal collection of data, this may become even more crucial, as the information will be increasingly "complete," potentially containing electroencephalographic (EEG) or physiological data alongside audio and video.

The major questions to ask oneself prior to recording or releasing affect data include, according to (Ragin & Amoroso, 2011) the moral principles that guide the research, how the ethical issues

influence the selection of the research problem itself, and the conduct of the research. These further include the responsibility one has toward the subjects as to whether they were sufficiently informed and whether one has obtained their consent. In fact, obtaining consent a priori may be challenging, because this might reduce the spontaneity and naturalness of the data. Ethical issues are also concerned with the question of which parts to release and/or publish. In multimodal data collection, subjects may, for example, agree to different levels of release for different contained modalities. An important question is further whether the research will be beneficial in some way in the near or far future to the subjects themselves. Finally, although it may seem obvious, it is important to point out that participants in studies should not be harmed in any way.

Recording and Reusing

The actual data are either obtained by recording new data or, as an often efficient alternative, by reusing existing material such as videos of political debates (Vinciarelli et al., 2009). The latter is, however, considerably more difficult if physiological measures are involved, as such data are usually only sparsely available—in particular in multimodal combination let alone in sufficiently emotional contexts.

The types of data obtained usually cover acted emotion, induced, for example, by recalling emotional memories or watching movie clips (see Chapter 32, this volume, on emotion elicitation methods for affective computing), and naturalistic emotion. However, in particular, the recording of multimodal data may make it difficult to record highly naturalistic data, as laboratory settings may be needed and may influence the recording. However, with the increasing availability of mobile and wearable ubiquitous devices, this challenge is expected to ease over time.

Collecting Metainformation

It may sound trivial, but in fact only a few databases contain rich metainformation on the subjects involved, the situational context, and so on. Besides obvious demographic aspects such as the age or gender of subjects, information such as cultural background, height, or spoken dialect may be beneficial in some cases. In particular, the personality of the subjects may be of interest in further data analysis. This can be assessed by standardized personality tests such as the 10-item questionnaire by Rammstedt and John (2007). For the different modalities, metainformation may in addition contain the recording equipment used and how it was synchronized.

Synchronizing Streams

One major issue during multimodal recording can be the synchronization of input streams. Even for the synchronization of audio and video, this may become a challenge if several microphones and cameras are involved. In other cases, such as the combination of worn physiological devices alongside video or depth capture, the recording may occur on individual devices that are not routed via the same computer. Aligned time stamps or markers are straightforward practical solutions allowing for later synchronization. Usually these markers may need to be repeated during a take (or trial) to compensate for temporal deviations. Another option, albeit usually involving greater effort and potentially leading to suboptimal solutions, is postrecording machine-based alignment, as by dynamic programming (Gunes, Piccardi & Pantic, 2008). However, this is generally an option only for the alignment of several captures from the same modality. Cross-modal alignment can become considerably more challenging.

Modeling

There are two decisions to be made in finding an appropriate model of emotion before (human) data labeling can start: (1) the emotion model and (2) the temporal unit of analysis. The choice of the most appropriate emotion model (e.g., continuous or categorical) can be influenced by the types of modalities involved in the multimodal setup. In any case, the temporal unit of analysis will usually be trickier to determine. For example, physiological measures and video could be annotated on a per-frame basis, whereas acoustic parameters are usually extracted over larger chunks, such as words or turns, and textual parameters may be most informative over whole phrases or dialogue acts. A recently adopted compromise is to choose a fully continuous annotation (Gunes & Schuller, 2012). *Fully* here means that the annotation should be continuous in its emotion dimension(s), such as arousal and valence, but also in time (i.e., an emotion value is given, for example, every 100 milliseconds). This allows for diverse mappings—for example, by averaging over a certain chunk. Another option is to use multiple models, which enriches the flexibility of the database but requires considerable extra effort. In the case of multimodal data, these different models could be applied for modality-specific annotation.

Labeling

Emotion labeling is probably the major effort besides the actual collection of the data. In a

multimodal database, this labeling can be tricky, as not all modalities can be easily annotated by a human rater. For example, physiological signals may be difficult to interpret. In fact, some recorded modalities may even serve as additional information for labeling, such as physiological measures in the case of arousal or video data for the annotation of physiological data.

Self-assessment is not always an option. Thus, usually several external labelers serve to approximate the "expertise of the mass" (e.g., by majority voting or by taking the mean and median in the case of continuous emotion models). The number of labelers should usually be increased with increasing subjectivity or ambiguity of the target labeling task at hand and the complexity of the chosen model.

Interestingly, multimodal recordings can be annotated modality-wise or in combination, which can lead to considerable differences. For example, acoustic and physiological data usually better convey arousal, whereas video or textual data are particularly well suited to convey valence (Karadogan & Larsen, 2012). Note, however, that not all modalities are necessarily present at all times. For example, speech is available only when a subject is talking.

Standardizing

A number of standards exist that may be considered to foster compatibility of the metadata and the annotations (Schröder et al., 2007). Of these, EmotionML (see Chapter 29, this volume) is a particular example of a markup language recommended by the World Wide Web Consortium (W3C). It allows for very high flexibility, including the use of one's own emotion lists. Further, it includes a number of mechanisms to describe multimodal data and to add contextual information. An earlier standard is EARL—the emotion annotation and representation language (Schröder, Pirker & Lamolle, 2006). It was partly suggested by the same researchers as EmotionML and laid the foundations for it.

Partitioning

Partitioning in the sense of dividing the affect data into partitions for different phases in modeling, optimizing, and testing is a crucial factor: If multimodal affect databases are not prepartitioned by their creators, one runs the risk that those working on the data will use diverse partitioning schemes, rendering the comparison of results and findings almost impossible. It is thus strongly recommended that a default or suggested form of partitioning be provided at the time of release.

Because data evaluation should ideally be based on test partitions that have not been "seen" during model creation and system optimization, *development* partitions are needed in addition to *training* and *test partitions*. A solution for using as much data as possible for all partitions is cross-validation, where the overall corpus is partitioned into J sets of equal size. These should be stratified (i.e., each set should show the same distribution of instances among classes or in the continuum in the case of numeric labels). For multimodal data, stratification may also aim at a good balance of presence and diversity in each modality in each partition. The evaluation is repeated J times, changing the roles of the partitions. Further criteria for partitioning include independence of subjects, context, and so on. A particularly frequent example is leaving out a subject or subject group at a time. Next, one wishes to keep good balance of all factors throughout the partitions. Further, partitioning should ideally be transparent and easy to reproduce. Thus random partitioning can be considered as a somewhat suboptimal choice, as one must provide the instance list or random seed and random function in order to allow for others to reproduce the partitions.

Verifying Perception and Baseline Results

An independent perception test with individuals other than the annotators may provide useful insights into the reliability of the annotation. Often, only a partition of the data such as the test set is used in such experiments, albeit with a potentially higher number of participants. This independent perception evaluation is then often used to compare a system's performance.

Again, as with the original labeling, such a study may be conducted individually per modality or for modality combinations. In addition, crowdsourcing (see Chapter 30, this volume) may be appropriate. A positive trend is to further include machine-based baseline recognition results in a data release for the orientation of others working on the data. These could give "just a rough first impression" on how difficult the task is for a machine and should ideally include results for unimodal and (different combinations of) multimodal fusion.

Releasing

The release of the data usually first requires the design of an end-user license agreement. Then, obviously, the highest spread and usage of the data

can usually be reached by making the data (almost) freely accessible. Ideally, the data will be accessible directly via the Internet, albeit restricted access must usually be guaranteed owing to the private nature of affect data. Another option is to release the data in the framework of a comparative or competitive evaluation campaign, such as the first two of their kind dealing with multimodal affect data—the Audio/Visual Emotion Challenges held in 2011 (Schuller et al., 2011a) and 2012 (Schuller et al., 2012a) (see also Chapter 18, this volume).

Quality Assessment—Is This Really Joyful?

Ground Truth Versus the "Gold Standard"

In affect computing, the "gold standard" is practically never reliable—that is, the training and testing labels themselves are ambiguous to a certain degree, as the emotion of a subject is usually difficult to assess, even in self-assessment (see Sneddon et al., 2012). Further, emotions are complex and often may not be mapped unambiguously to a single category or point in space (Schuller et al., 2010). The terms *ground truth* and *gold standard* are often used more or less synonymously in the literature; here, we want to define *ground truth* as the actual truth as measured "on the ground"—the term itself in fact originated in the fields of aerial photographs and satellite imagery—as compared with the gold standard that might ideally be identical to the ground truth; however, it might also be the (slightly) error-prone labeling as seen from the "sky above."

In interpreting results, one thus has to bear in mind that the reference is usually the gold standard and not necessarily the ground truth. This has a double impact: On one hand, trained models of computer systems that process affect data are error-prone. On the other hand, the test results have to be taken with a grain of salt, given that a "classification error" might not be so wrong in ambiguous cases. Thus, in order to achieve a reliable gold standard close to the ground truth, usually several annotators are used. This method also offers interesting implications for machine learning, as systems can be trained on individual annotator tracks in addition to an overall gold standard.

Measuring Reliability—From Alpha to Kappa

There are several commonly used measures to assess agreement among the labelers—the *interrater reliability*—in the usual case where not a single rater but around 4 to 10 or more raters are involved. If the affect is modeled continuously, the (mean) correlation coefficient (CC) or (average) mean linear/absolute error (MLE, MAE), mean square error (MSE), and standard deviation among labelers are frequently used. If only one measure is to be considered, it may be the correlation, as it is usually more informative in the given case of a gold standard without a reliable reference point (Schuller et al., 2012a).

In the case of categorical modeling, a variety of measures can be employed for agreement evaluation, such as Krippendorff's alpha or Cohen's or Fleiss' kappa. As a continuum can be discretized, the latter statistics can also be used in this case—often with a linear or quadratic weighting. Pearson's intraclass correlation coefficient and Spearman's rank correlation coefficient rho are particularly suited for such ranked intervals—albeit only for two raters. Fleiss' Kappa K—a generalization of Scott's pi for more than two raters—is one of the most frequently encountered measures in the field. It requires all raters to rate all data. If labelers agree throughout, K equals 1. If they agree only on the same level as chance would, then K equals 0. Negative values indicate systematic disagreement. According to Landis and Koch (1977), values of.4 to.6 indicate moderate agreement, and values above are considered good to excellent agreement. However, these levels of reliability are difficult to achieve with affecte labeling given the often ambiguous and partially subjective nature of affect data.

Weighting Evaluators—I Don't Trust This Labeler

Further, if some labellers provide a rather different annotation than the majority do, labelers can be weighted individually in order to reach a more consistent gold standard. The justification is that some labelers may lack concentration if they have to label huge amounts of data or do not take labeling seriously at all times. This may become particularly relevant if naïve labelers in large number are involved, as by crowdsourcing via the Internet.

The evaluator-weighted estimator (EWE) as described by Grimm & Kroschel (2005) provides an elegant model to reach a rater-weighted gold standard. EWE's average of the individual evaluators' responses takes into account the fact that each evaluator is subject to an individual amount of disturbance during evaluation. The weights measure the correlation between the individual annotator's estimations and the average ratings of all evaluators.

If the weights are constant among raters, the gold standard is the simple mean of the raters' continuous labels.

An alternative can be to filter outliers; for example, by Peirce's outlier detection (Karadogan & Larsen, 2012). One can also imagine combinations of generally weighting raters and filtering or weighting individual labels.

Efficiency—How to Save Annotation Labor

Apart from the actual collection of data, the annotation usually consumes the most resources. In this section, five strategies that are recently gaining interest in the community are presented. These strategies are used to reduce costs and "feed" machines with partly self-labeled data, to reuse resources by pooling existing databases, or to avoid risking the loss of potentially interesting data by machine-based preselection of the "interesting bits."

Active Learning: "Help me, I'm a machine"

Active learning (Zhang & Schuller, 2012) aims at finding a needle in a haystack when massive amounts of data are available of which only a few items are of interest. In affect data, the haystack is usually the neutral data around the origin in a dimensional model, and the needle is usually "non-neutral" data, such as anger. Rather than having the human look over all recorded data, the machine tries to locate potential cases of interest and then asks the human for help or confirmation. As a supervised learning approach, active learning thus aims to minimize the amount of human supervision required in cases where one can afford to lose samples of emotional data; the goal thus is to identify the "most informative" samples in the unlabelled data—that is, those that we would gain most by if they were manually labeled—and then to present only these sample to human labelers. Several approaches have been investigated for selecting these most informative samples (Settles, 2010). A well-known method is *uncertainty-based* active learning, in which the active learner determines the certainties of the predictions on the unlabeled data based on posterior probabilities. The samples with least certainty are then generally presented to the labelers for annotation. Another common strategy is a *committee-based* method. Predictions for unlabeled data are made by multiple classifiers. The samples considered as most informative are those with the lowest agreement. Other methods include the *expected-error-reduction* method, which aims to measure how much the generalization error is likely to be reduced; the *expected-model-change*–based method, which chooses the instances that impact the current model the most; and the *diversity-density*–related method.

A major drawback of these methods is that they ignore the problem of class unbalance or the issue of scarcity of certain classes, as is mostly the case in affect data. Zhang and Schuller (2012) thus present a tailored sparse-instances–based strategy that selects the samples "likely to be" the fewest to be annotated manually from the candidate data in the pool.

In the case of multimodal data, different modalities may be used to search for different aspects of interest in the data. For example, physiological measurement may indicate moments of high arousal. Then, preferably at these moments, the human may be asked to annotate the data from an audiovisual point of view.

Semisupervised Learning: "Okay, I can label this!"

One step further is, after a supervised initialization, allowing the machine to label data by itself without further supervision—the so-called *semisupervised approach* (Zhang et al., 2011).

Assuming sufficiently robust automatic emotion recognition engines, unlabeled data can be classified and integrated into an iterative retraining process. Two parameters are then of primary interest: the *iteration number* indicating how often the unlabeled data are relabeled by the incremental addition of new data, and the *upsampling factor*, which can be used to weight the original human-labeled data more strongly than the later added machine-labeled data.

As a rule of thumb, roughly ten times the amount of unlabeled data are needed in comparison to labeled data in order to obtain the same gain as when human-labeled data are used exclusively. This makes it clear that this method is particularly suited ton cases where practically infinite amounts of data are available. An example can be the exploitation of audiovisual resources on the web, as on YouTube.

So far, initial studies in semisupervised learning for emotion recognition in speech show promising results (Mahdhaoui & Chetouani, 2009), and it will be interesting to see how far self-learning affect processing can take us for multimodal data. It seems worth mentioning that first successes are reported to entirely synthesize affect data for usage in recognition systems (Schuller et al., 2012b). This renders

the labeling need completely obsolete. In multimodal tasks, this may be even more challenging, but with the increasing availability of multimodal affective agents, the possibility of using their synthesized affective behavior exists to train other systems.

Unsupervised Learning: "Trust me—I'm a machine"

Without any human-labeled data, *unsupervised learning* needs to find its own categorization, as by the EM algorithm, k-means, or other techniques (Wöllmer et al., 2009). Unsupervised learning may be of interest, as it omits the need to find an appropriate model. In most clustering approaches, however, the number of target clusters, (i.e., classes) needs to be given as input. This can be used as a design parameter to keep this number either low to aim at a rough yet robust categorization, or to keep it high in case of a need for a fine-grained model. In multimodal data, the clustering can first be carried out individually per modality to better take peculiarities into account, such as the previously mentioned higher correlation of acoustics with arousal and video feed with valence, and so on.

Generally speaking, fully unsupervised learning can be particularly useful in autonomous systems that need to recognize and synthesize affect—they use automatically learned emotional clusters in analysis and the appropriate counterpart for synthesis. It must be ensured, however, that the derived clusters are not too strongly influenced by other factors, such as the gender of the recorded subjects or similar differences. If such an influence gains too strong an effect, the data may first be separated according to these factors.

If human feedback or system context can be exploited during runtime, learning can further be *reinforced* (Hyung-il Ahn, 2010). This may lead to genuine online "lifelong learning" of affective systems that recognize and react to affect and, from the reaction of their users and the environment, improve their future processing abilities. Such systems can then collect data themselves "in the wild." This may become particularly challenging if the *emotional* reaction itself is used as contextual knowledge.

Shared Learning: "Together we're best"

In fact, active learning and semisupervised learning attack the same problem from opposite directions (Settles, 2010): Semisupervised learning exploits the learner's assumptions on the unlabeled data; active learning aims to explore the unknown aspects of the data. This lets one strive to combine the two strategies. Figuratively speaking, it means shared work between the (pretrained) machine and the human during annotation, with the machine having the lead: It labels data by itself if it is confident that it can label correctly. From the remaining unlabeled data, it decides which cases appear to be interesting and should be labeled by humans while disregarding the rest. Again, this may benefit from modality-specific analysis, so that the machine prefers some modalities such as audio and video for human feedback but will be more "aggressive" in labeling other modalities by itself (e.g., when it is labeling physiological data). Further, the machine could ask for human help if it came to different conclusions in looking at different modalities.

Pooling Data: "Let's save the environment"

The reusage of existing labeled data seems to be straightforward in general. However, in the field of affective computing, this is less obvious, as data often come labeled in different models, as in categories or dimensions or chunked in different units of time. Therefore to obtain a larger pool of labeled data by reusing and uniting existing material, a mapping scheme may be needed. The dimensional model offers an elegant solution in this case, as categories can be mapped onto coordinates and corpora with different label sets can thus be united based on the dimensional model. However, this mapping step has to be carried out by experts (Schuller et al., 2011b) or optimized (e.g., by machine learning strategies). In addition, algorithms can then select and weight instances. For example, the joyful data of a particular database may fit less well than the angry data of the same set. Further, some databases may be better suited for a certain target domain, but further data still enriches the pool. In such a case, these instances may be weighted (e.g., by repeated upsampling). It seems particularly interesting to find measures to compare corpus and emotion data similarity a priori—that is, without the need for computationally expensive repeated model training (Brendel et al., 2010).

As for temporal unification, usually the larger unit of time (e.g., word or turn in speech analysis) needs to be taken as basis, since emotion may only be *quasistationary* or not at all stationary along the larger unit; think, for example, of a phrase like *"Thursdays are quite ok, but I hate Mondays!"* The overall phrase may be labeled with negative valence, whereas the first frames are probably not negative

(Batliner et al., 2010). For example, the average over frames can be used to map this shorter unit of annotation onto a longer one, such as the named words or turns.

Existing Multimodal Resources—What's There?

Luckily, a large number of multimodal resources of affect data currently exist. However, only a selection of representative corpora can be presented in this chapter. One can also expect further corpora to be available in the near future.

Exemplary Audiovisual Resources

In the following, some characteristic resources are introduced, of which some are immediately available for download; others may be per request.

Some of the first multimodal databases were of bimodal nature. Most prominent in this group are audiovisual databases, which to the present day form by far the lion's share of multimodal affect databases. As these feature speech, some tend to describe them as more than bimodal, given that acoustic and textual cues can be interpreted with different means. However, in the strict sense of modality, these can be subsumed. In a similar fashion, some consider partly contained motion capture information from video as an additional modality, which can, however, be subsumed under the video modality.

An example of single-person noninteractive data is the freely accessible eNTERFACE corpus (Martin et al., 2006). This corpus targets the Ekman "big six" basic emotions by short story–based induction for around 40 subjects. The stories are basically short texts that are read by the subjects before enacting given target phrases that fit the context of these short texts.

Next, the "mind reading" corpus by Baron-Cohen and Tead (2003) features an extremely high diversity of more than 400 emotional nuances performed by six professional actors.

A particularly early example of human-technology recordings is the SMARTKOM database of Wizard of Oz–type human-machine interactions with an information service in public, home, or mobile environment, including affect annotation. The users are interacting with information services in rather natural ways—the functionality was simulated by human operators. The database features a categorical annotation in nine categories and 224 subjects (Schiel, Steininger, & Türk, 2002).

The freely available SEMAINE database (McKeown, 2012) deals with high-quality audiovisual recordings of 150 participants. In these recordings, humans interact with four different versions of a sensitive artificial listener agent. It includes 959 conversations (approximately 5 minutes each). Six to eight raters labeled the data in five dimensions. A subset of these data was used in the two Audio/Visual Emotion Challenges (Schuller et al, 2012).

Aiming at more naturalistic data, the EmoTV corpus (Abrilian et al., 2005) exploits film clips. Another example of the emerging trend to record multiple-subject dyadic interactions, rather than human-computer or human-technology interaction, is The TUM Audio Visual Interest Corpus (TUM AVIC) (Schuller et al., 2007). It features audiovisual recordings of conversations between a product presenter and 21 participating subjects. Transcriptions including nonverbal outbursts are available as metainformation, and this data set provides a one-dimensional labeling by averaging over four labelers. Its audio track was featured in the INTERSPEECH 2010 Paralinguistic Challenge's Affect Sub-Challenge.

Further, the IEMOCAP database (Busso et al., 2008) contains 12 hours of acted multimodal recordings of multiple subjects. Besides video and speech, text transcriptions and facial motion capture are provided. In the dyadic sessions, actors perform improvisations or scripted scenarios. These were selected to elicit emotional expressions. The database is annotated by multiple annotators and contains categorical labels, such as anger, happiness, sadness, neutrality, and continuous emotion primitives including valence, activation, and dominance.

A more recent example of dyadic interaction is the UCS CreativeIT database. Its purpose is to study affective communication and interaction between humans (Metallinou et al., 2010). The data contained are based on improvisation of pairs of theater actors. These were recorded with cameras; motion capture markers were placed over their full bodies, and close-talking microphones were worn by the actors to capture their speech. The rather long, unsegmented recordings last from 2 to 8 minutes. The annotation was carried out fully continuously in value and time for the dimensions of activation, valence, and dominance.

Finally, a multimodal database particularly designed for mimicry analysis has been introduced by Sun et al. (2011). Eighteen synchronized audio and video sensors were used in the recording setup and two dyadic interaction settings were given.

In these, participants engaged in a discussion on a political topic as well as in a role-playing game. Overall, the database contains 54 recordings from 40 participants and 3 confederates. Metadata such as dialogue acts or turn taking are released together with the recordings.

Exemplary Resources Containing EEG and Physiological Data

More recently, increasingly multimodal data with additional multiple modalities have appeared. Three examples to illustrate the available resources are discussed here.

The QMUL-UT EEG dataset features multimodal affect data of 17 subjects, including EEG and physiological signals such as electrooculography (EOG), galvanic skin response (GSR), heart rate, respiration, and temperature. The subjects were watching seven video sequences for each of seven categories depicting events, followed by either a matching or a nonmatching emotion label (Koelstra, Muehl & Patras, 2009).

The freely available database for emotion analysis using physiological signals (DEAP) (Koelstra et al., 2012) contains spontaneous physiological signal recordings and face videos (not for all participants) of 32 participants watching and rating online their emotional response to 40 music videos along the scales of arousal, dominance, and valence as well as ratings of how much they liked and were familiar with the videos. The authors also provide classification results using various features (from the EEG, peripheral physiological signals, and other modalities) and combinations of features, also performing single-trial (single-participant) classification (for the scales of arousal, valence, and liking). They report that modalities appear to perform in a moderately complementary fashion, where EEG performs best for arousal and peripheral physiological signals for valence.

Next, Ringeval et al. (2013) collected the freely available French RECOLA multimodal corpus of remote collaborative and affective interactions. Its 46 subjects were recorded in dyads during a video conference for collaborative task completion. Modalities comprised audio, video, electrocardiogram (ECG), and electrodermal activity(EDA). Six annotators rated continuous arousal and valence, as well as social behavior labels on five dimensions. Further, self-report measures are included.

Finally, the freely accessible multimodal affect database for affect recognition and implicit tagging (MAHNOB-HCI) (Soleymani et al., 2012) is a collection of various modalities recorded in a synchronized manner. The recorded cues include six camera views of the face and head, sound from both a head-worn microphone and one located in the room, eye gaze, pupil size, and peripheral/central nervous system physiological signals: ECG, EEG, GSR, respiration amplitude, and skin temperature. The authors also provide baseline emotion recognition using three modalities and implicit labeling results for two modalities. The 30 participating subjects watched 20 emotional videos and subsequently self-reported their felt emotions. They also judged videos or images with or without correct or incorrect emotion labels.

As indicated above, with the increasing availability of wireless sensors, more databases with data from outside the lab and with physiological information can be expected to be available in the near future.

Conclusions and Future Avenues—Wrapping Up

The above discussion shows that multimodal affect data are increasingly available. However, more data recorded in real-life situations (Lucey et al., 2011) outside of the lab will be needed and will likely be seen soon. Ideally these will also feature a higher diversity of participants, labeling, contained cultural aspects, languages, and situational context. Luckily, machine intelligence can help to reduce human labor costs during the annotation of data by active learning, identifying the interesting data instances for annotation in large unlabeled collections, or by weakly supervised learning, having the machine label when it is sufficiently confident it "can do the job."

Future initiatives could help foster combined community efforts for merging and common labeling of resources and could also make such desperately needed larger amounts of resources accessible with prepartitioning, human perception results, baselines, and rich metainformation.

References

Abrilian, S., Devillers, L., Buisine, S., & Martin, J. (2005). EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *Proceedings International Conference on Human-Computer Interaction.*, Las Vegas, NV. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ahn, H. (2010). *Modeling and analysis of affective influences on human experience, prediction, decision making, and behavior.* PhD thesis. Cambridge, MA: Massachusetts Institute of Technology.

Baron-Cohen, S. & Tead, T. (2003). *Mind reading: The interactive guide to emotion*. London: Jessica Kingsley.

Batliner, A., Seppi, D., Steidl, S., & Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. In *Advances in human computer interaction, special issue on emotion-aware natural interaction*, Article ID 782802. Hindawi.

Brendel, M., Zaccarelli, R., Schuller, B., & Devillers, L. (2010). Towards measuring similarity between emotional corpora. In *Proceedings LREC 2010* (pp. 58–64). Valletta, Malta: ELRA.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S.,...Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database, *Journal of Language Resources and Evaluation, 42* (4), 335–359.

Grimm, M., & Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. In *Proceedings of the ASRU 2005* (pp. 381–385). Cancun, Mexico: IEEE.

Gunes, H., Piccardi, M., & Pantic, M. (2008). From the Lab to the real world: Affect recognition using multiple cues and modalities. In *Affective computing: Focus on emotion expression, synthesis, and recognition.*, (pp. 185–218). Vienna, Austria: Tech Education and Publishing.

Gunes, H., & Schuller, B. (2012). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing, 31* (2), 120–136.

Karadogan, S., & Larsen, J. (2012). Combining semantics and acoustic features for valence and arousal recognition of speech. In *Proceedings of the 3rd International Workshop on Cognitive Information Processing.*, Parador de Baiona, Spain. IEEE.

Koelstra, S., Mühl, C., & Patras, I. (2009). EEG analysis for implicit tagging of video data. In *Proceedings ACII, Affective Brain-Computer Interfaces Workshop* (pp. 27–32), Amsterdam, The Netherlands.

Koelstra, S., Mühl, C., Soleymani, M., Yazdani, A., Lee, J.-S., Ebrahimi, T.,...Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing, 3* (1), 18–31.

Labov, W. (1984). Field methods of the project in linguistic change and variation. In *Language in use* (pp. 28–53). Englewood Cliffs, NJ: Prentice-Hall.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., & Matthews, I. (2011). Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Proceedings IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 57–64), Santa Barbara, CA.

Mahdhaoui, A., & Chetouani, M. (2009). A new approach for motherese detection using a semi-supervised algorithm. In *Proceedings IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, pp. 1–6.

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The enterface'05 audiovisual emotion database. In *Proceedings International Conference on Data Engineering* (pp. 1–8)., Atlanta, GA. IEEE

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroeder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing, 3* (1), 5–17.

Metallinou, A., Lee, C.-C., Busso, C., Carnicke, S., & Narayanan, S. (2010). The USC CreativeIT database: A multimodal database of theatrical improvisation. In *Proceedings LREC Workshop on Multimodal Corpora.*: ELRA.

Ragin, C. C., & Amoroso, L. M. (2011). The ethics of social research. In *Constructing social research*, 2nd ed. (pp. 59–89). Thousand Oaks, CA: Sage.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*, 203–212.

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*. IEEE Face & Gestures 2013, Shanghai, China).

Schiel, F., Steininger, S., & Türk U. (2002). The SmartKom Multimodal Corpus at BAS. In *Proceedings LREC 2002* (pp. 200–206), Las Palmas, Gran Canaria, Spain, ELRA.

Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C.,...Wilson, I. (2007). What should a generic emotion markup language be able to represent? In *Proceedings ACII 2007* (pp. 440–451). Berlin and Heidelberg: Springer.

Schröder, M., Pirker, H., & Lamolle, M. (2006). First suggestions for an emotion annotation and representation language. In *Proceedings LREC 2006 Workshop on Emotion: Corpora for Research on Emotion and Affect*, Genoa, Italy. ELRA.

Schuller, B., Douglas-Cowie, E., & Batliner, A (2012). Guest editorial: Special section on naturalistic affect resources for system building and evaluation. *IEEE Transactions on Affective Computing, 3* (1), 3–4.

Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H., & Rigoll, G. (2007). Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of the ICMI 2007* (pp. 30–37), Nagoya, Japan. ACM, ACM.

Schuller, B., Valstar, M., Cowie, R., & Pantic, M. (2012a). AVEC 2012—The continuous audio/visual emotion challenge. In *Proceedings of the ICMI 2012.*, Santa Monica, CA. ACM, ACM.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M. (2011a). AVEC 2011—The first international audio/visual emotion challenge. In *Proceedings ACII 2011* (Vol. II, pp. 415–424)., Memphis, TN. Berlin, Heidelberg NY: Springer

Schuller, B., Zaccarelli, R., Rollet, N., & Devillers, L. (2010). CINEMO—A French spoken language resource for complex emotions: Facts and baselines. In *Proceedings of the LREC 2010* (pp. 1643–1647). Valletta, Malta. ELRA.

Schuller, B., Zhang, Z., Weninger, F., & Burkhardt, F. (2012b). synthesized speech for model training in cross-corpus recognition of human emotion. *International Journal of Speech Technology, 15* (3), 32–41.

Schuller, B., Zhang, Z., Weninger, F., & Rigoll, G. (2011b). Using multiple databases for training in emotion recognition: To unite or to vote? In: *Proceedings of INTERSPEECH 2011* (pp. 1553–1556)., Florence, Italy. ISCA.

Settles, B. (2010). Active learning literature survey. In *Computer sciences technical report 1648*. Madison: University of Wisconsin-Madison.

Sneddon, I., McRorie, M., McKeown, G., & Hanratty, J. (2012). The Belfast induced natural emotion database. *IEEE Transactions on Affective Computing, 3* (1), 32–41.

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3* (1), 42–55.

Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., & Pantic, M. (2011). A multimodal database for mimicry analysis. In *Proceedings ACII* (pp. 367–376)., Memphis, TN. Berlin, Heidelberg, NY: Springer.

Vinciarelli, A., Dielmann, A., Favre, S., & Salamin, H. (2009). Canal9: A Database of Political Debates for Analysis of Social Interactions. In: *Proceedings ACII*, Amsterdam, The Netherlands, IEEE/Humaine Association.

Wöllmer, M., Eyben, F., Schuller, B., Douglas-Cowie, E., & Cowie, R. (2009). Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks. In *Proceedings INTERSPEECH 2009* (pp. 1595–1598). Brighton, UK. ISCA, ISCA.

Zhang, Z., Weninger, F., Wollmer, M., & Schuller, B. (2011). Unsupervised Learning in cross-corpus acoustic emotion recognition. In *Proceedings of the ASRU 2011*(pp. 523–528). Big Island, HY. IEEE.

Zhang, Z., & Schuller, B. (2012). Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *Proceedings INTERSPEECH 2012*, Portland, OR. ISCA, ISCA.