Perceptual Information Fusion in Humans and Machines

Albert Ali Salah

Centrum voor Wiskunde en Informatica,
Amsterdam, The Netherlands
a.a.salah@cwi.nl

ABSTRACT

Humans perceive the world through different perceptual modalities, which are processed in the brain by modality-specific areas and structures. However, there also exist multimodal neurons and areas, specialized in integrating perceptual information to enhance or suppress brain response.

The particular way the human brain fuses crossmodal (or multimodal) perceptual information manifests itself first in behavioural studies. These crossmodal interactions are widely explored in some modalities, especially for auditory and visual input, and less explored for other modalities, like taste and olfaction, yet it is known that these effects can occur with any two modalities.

The integration of sensory data is an important research area in computer science, and stands to benefit from the studies into the brain function; many biological processes serve as models for computer algorithms. On the other hand, computer models of sensor integration are built on mathematical principles, and provide normative insights into the functioning of the brain.

This paper surveys the psychological and neurological findings pertaining to human multi-sensor fusion, followed by a brief review of the relevant computer science terminology and modeling approaches. The potential of an interdisciplinary approach to information fusion encompassing neuroscience, psychology and computer science have recently been recognized, and a multidisciplinary workshop on biologically inspired information fusion was organized to bring researchers together to determine a common agenda. The conclusion summarizes the agenda of research outlined at the workshop and attempts to raise research questions for the future.

INTRODUCTION

Neurologists and psychologists have been working on individual senses to understand how perception occurs in humans. Today, it is accepted that perception is not dependent on a single modality at any given time, but has a multimodal nature; perceptual information from different senses can be fused to produce a percept manifestly different than the individual sensations that constitute it. Thus, examining the individual senses is not enough to understand perception, the interaction and the dynamics of multimodal fusion must be studied.

Computers are not very successful in perceptual tasks so naturally and easily performed by humans in everyday life. Yet, information fusion is a familiar concept for computer scientists working in the field of machine learning and pattern recognition. When available, fusing different data channels and different representations can bring robustness and increased accuracy to learning algorithms. The mathematical and theoretical models developed in this field allow for a two-way information flow between computer science and cognitive sciences. On one hand we would like to express the fusion processes in the brain in a principled, mathematical way, which in turn would allow us to use computer simulations to gain more insight into these mechanisms. On the other hand understanding the brain mechanisms may help us in building better computational methods for crossmodal fusion, improve sensor technologies and open up new application areas (Murphy, 1996).

PERCEPTUAL FUSION

The sensations created in our environments rarely exist in a single modality. Especially visual events are usually accompanied by auditory events. When an object falls and breaks, the simultaneous reception of the visual and auditory inputs form a single perception of this event. Under certain conditions, the human brain perceives sensations from different modalities as a whole. This property can be seen as a natural consequence of the evolutionary process. The natural environment consistently produces multimodal information, and humans have evolved optimized sensory organs to perceive the environment as well as a brain to represent and reflect the multimodal nature of this information. Hence, we can treat the unconscious sensory fusion process as a problem automatically solved by the brain, and model it with the help of computers.

Simultaneous perceptual information is often fused pre-consciously in the brain. Principles that resemble Gestalt rules play a role in the combination of information across senses. The

process of combining two senses into a single percept is called *pairing* (Bertelson, 1999). However, fusion is not only pre-conscious, it happens in different levels, and conscious semantic information can influence fusion.

Experimental results have shown that two modalities need not be in perfect temporal synchrony for fusion. If there is a fixed spatial or temporal distance between two percepts, and the percepts show structural similarities across these distances, this may lead to fusion as well. We can classify the mutual interaction of perceptual modalities in three groups:

- *Perceptual fusion*: Perceived modalities are fused into a new and novel perception.

- *Immediate crossmodal bias*: The information coming from one sensory channel influences and changes the perception of information from another sensory channel. This influence can be a strengthening or a weakening of the perception, or guidance of attention in one modality through cues in the other modality (Driver and Spence, 1998).

- *After-effects*: Perception in one modality has a lasting after-effect that influences later information processing in another modality. The time difference between these perceptions need not be small.

Perhaps the most important characteristic of perceptual fusion is that it is an automatic process, and takes place robustly, even in the absence of synchronization.

BEHAVIOURAL EVIDENCE

The understanding of the full complexity of human perceptual capabilities came with early behavioural studies. Researchers were working on perception since 1960s, designing experiments to probe which perceptual modalities were fused under what conditions, and the dominance and transfer relations between modalities. The advances in brain imaging techniques later allowed the researchers to re-run these experiments, this time observing brain activity in addition to behavioural evidence. In this section we will summarize the main experimental results obtained in these behavioural studies.

One of the earliest experimental settings to understand fusion involves a prism to shift visual perception. In these experiments the visual input conflicts with information from other modalities, as it does not reflect the true position of the event anymore. For instance a subject that receives a needle prick on one of his or her hands while watching it through the prism has two conflicting sensations on the location of the hand. As a result of this, the perception deviates from the true location in the direction of the misleading sensation. In this experiment,

the consistency in the timing, direction and speed of the action for both modalities is enough for perceptual fusion.

In the prism-based experiments, researchers have observed that subjects reaching for objects initially made predictable mistakes, as the prism shifts the visual perception, but later corrected these mistakes as the new visual orientation is habituated. Once the prism was removed, the subjects did not immediately revert to their ordinary visual mode, but made errors in the other direction. These experiments are demonstrative of after effects between vision and proprioceptive perception. Even though the initial experiments indicated an apparent dominance of vision to proprioceptive perception, later experiments do not report a significant dominance in either way (Hay, Pick and Ikeda, 1965; Bertelson and de Gelder, 2004).

Experiments on animals also provide interesting findings. Knudsen and Knudsen have applied the prism paradigm to young barn owls, and observed that after some habituation period the auditory perception shifts in consistence with the shifted visual perception (Knudsen and Knudsen, 1985). It took more than a month for the auditory perception to become completely consistent with the visual perception. Once the prism was lifted, the visual habituation is fast, but auditory perception required several weeks to return to its original state (Knudsen and Brainard, 1991). Another interesting experiment from the same group demonstrates that owls that are raised with one ear plugged make systematic errors in auditory localization, but correct these mistakes in a couple of weeks after the plug is removed (Knudsen and Knudsen, 1985). However, owls could not correct these mistakes if they were made to wear blindfolds, preventing them access to visual information. These experiments demonstrated the existence of influential neural correlates that link auditory perception to vision, and identified crossmodal influences relevant to the development of the perceptual system.

In multimodal fusion, dominance of one modality over the other is frequently observed, depending on the context. In humans, vision is in general the more dominant modality (Rock and Harris, 1967). This phenomenon may have statistical grounds, or it may be related to the fact that the visual cortex is physically more dominant in the brain when compared to other modalities. These two reasons are related through an evolutionary argument. Natural selection results in brains that have more resources for more informative modalities, i.e. vision. Behavioral studies of dominance can be supported with neurological findings and computer simulations based on these arguments.

In the remainder of this section, we will review evidence of fusion between pairs of modalities. Experiments on perceptual transfer and dominance relations have particularly focused on audio and vision, and this is where we will start.

Audio and Vision

In a paper published in 1976, McGurk and MacDonald have shown that involuntary and consistent multimodal fusion takes place for synchronized auditory and visual perception (McGurk and MacDonald, 1976). Their experiments were based on observing the perception of subjects in the face of auditory, visual and auditory+visual information. With the help of a video recording, the subjects have seen a woman's face mouthing the syllable "ga", which is the visual input. The synchronized auditory input is a woman's voice, saying the syllable "ba". The subjects who received only the visual input perceived the syllable "ga", and the subjects who received only the auditory input perceived the syllable "ba". However, subjects who received both perceived the syllable "da", which is indicative of multimodal fusion producing a completely different perception. This experiment has been repeated many times, in many forms, and the effect has been demonstrated with people speaking different mother tongues (Massaro, Tsuzaki, Cohen, Gesi and Heredia, 1993), with children (Rosenblum, Schmuckler and Johnson, 1997), and even in cases when the subject is not aware of looking at a face (Fowler and Dekle, 1991).

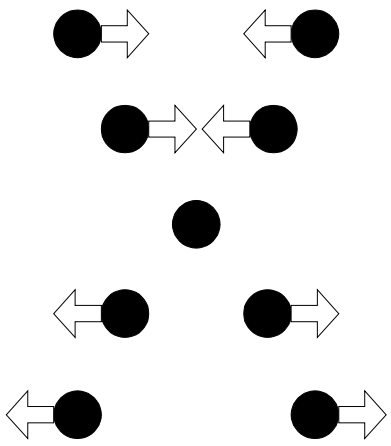

Figure 1 Moving disks experiment. Two disks move towards each other, meet at a point, and continue their motion in opposing directions.

An experiment in which auditory information changes the perception of visual information is shown in Figure 1. In this experiment, two disks that move towards each other meet at one point and continue in opposing directions. Without an accompanying sound, most subjects

perceive that disks just pass through each other to continue their way (Sekuler, Sekuler and Lau, 1997). If, however, a 2.5 millisecond long sound at 75 dB is played at the moment of their overlapping, most subjects perceived that the disks hit each other and bounce. Even when the synchronization of sound and image are disturbed, the perception of bouncing disks prevails. The auditory cues clearly change the visual perception.

An extensively studied crossmodal effect is the shift in the perceived location of a sound based on visual input. The ventriloquist effect demonstrates this phenomenon, as the mouth movements of the ventriloquist's puppet serve as a visual input to change the perceived location of the sound, shifting it form the ventriloquist to the puppet (Radeau and Bertelson, 1976). Recently, researchers have contrasted fMRI readings during the ventriloquist illusion with cases where auditory and visual cues are present without crossmodal fusion, and observed increased activity in the insula, superior temporal sulcus and parietal-occipital sulcus during fusion (Bischoff et al., 2006). In Calvert et al. (1997), it was shown that the auditory cortex is activated during lip-reading. These experiments point out to different neural structures linking these modalities.

In a series of experiments, Giard and Perronet established that: 1) perception of objects is faster for cases where visual input is accompanied by related auditory input, 2) the information fusion has an offset of 40 ms., 3) it has a distributed nature with respect to its neural involvement, 4) it has a flexible and adaptive nature, as the effects changed depending on the dominant modality (Giard and Perronet, 1999). They have used a technique based on event related potentials (ERP), which is frequently used for studies in crossmodal fusion (Stein and Meredith, 1993). In this technique, it is assumed that ERP recordings reflect activity in the perceptual processing areas. The activity levels during visual-only perception (V) and auditory-only perception (A) are summed up, and contrasted with activity levels obtained during multi-modal perception (AV).

In another ERP-based work, Talsma and Woldorff compared multimodal perception with auditory and visual perception (2005). In their experiments subjects were asked to focus their attention on a single-modality cue that occurs infrequently. During perceptual fusion, the activity levels were observed to be different than the total activity obtained by separate single-mode activities, and much higher for the target cue. Their results show that multi-modal fusion can occur in different levels of processing, can be influenced by attention, and is not necessarily automatic in all cases.

Shams, Kamitani and Shimojo (2000) have demonstrated how visual perception can be changed via auditory cues with an interesting experiment, where subjects seeing a flash of light and hearing multiple beeps at the same time perceived multiple flashes. This effect is robust to disturbances in the synchronization of light and sound, as well as to changes in the luminosity, colour and shape of the flash (Shams, Kamitani and Shimojo, 2002). The dominance of auditory cues is especially noteworthy, since visual input is usually more dominant (e.g. McGurk effect).

The results of this experiment are confirmed by Andersen, Tiipana and Sams (2004), who also investigated whether a change in the modalities would produce a similar effect, i.e. a single beep with multiple flashes producing a perception of multiple beeps. The experiments show that this effect is very difficult to obtain, and it is only possible when the sound levels are close to the hearing limits of the human ear. In an attempt for a unified account of dominance, the experimenters then contrasted four hypotheses to account for their findings:

1. *Discontinuity Hypothesis* (Shams, Kamitani and Shimojo, 2002): The modality in which the continuity is disrupted is more dominant.
2. *Modality Appropriateness Hypothesis* (Welch and Warren, 1980): The modality that is related to the aims of the subject is more dominant. For instance, if the subject is instructed to count the lights and sounds, auditory perception should be more dominant, as it is easier to distinguish and count separate instances by hearing.
3. *Information Reliability Hypothesis* (Schwartz, Robert-Ribes and Escudier, 1998): The modality with more reliable information is more dominant. Since vision is usually the dominant modality, it will be more dominant in the majority of tasks.
4. *Directed Attention Hypothesis* (Welch and Warren, 1980): The modality that receives attention is more dominant.

Andersen, Tiipana and Sams note that the results of their experiments are consistent with all hypotheses, except for modality appropriateness. They have explicitly observed that when the continuity of the information flow in a modality is disrupted, or if the subjects are instructed to pay particular attention to a modality, that modality is more dominant. We can argue that a more general attention-based hypothesis can account for both cases, as the former is a bottom-up attentional cue for the modality, whereas the latter is a top-down influence.

People can focus their attention on one speech thread in a crowded environment, and follow that thread, disregarding everything else. In the signal processing literature, this is called “the

cocktail party effect". It is difficult to disentangle different streams of speech by processing them on a computer; visual cues help people to achieve this task. In an experiment that demonstrates visual effects on audition with a cocktail party setting, Driver asked his subjects to listen to two superposed streams of speech, but to follow only one of them (Driver, 1996). The subjects were simultaneously shown a face on a screen, mouthing the speech the subject is supposed to follow. With the help of information fusion in the visual and auditory modalities, the subjects could distinguish the two speech streams much better in the presence of visual input. This experiment had another, non-trivial result: Based on previous crossmodal fusion research, we can predict that the visual cues will shift the perceived location of the speaker (for which there is crossmodal fusion) towards the screen. Indeed, it is shown that moving the screen away from the subject will move the perceived location away as well, thereby leading to a better separation of speech signals in comparison to the case where the screen is placed closer to the subject.

Fusion in other modalities

Crossmodal fusion can be observed in other perceptual modalities as well, but there are fewer published research results. This section presents several examples. Crossmodal influences between all possible perceptual channels suggest the possibility of a central perceptual fusion mechanism. Researchers have focused on common points of fusion in different modalities to investigate this hypothesis.

The sense of touch has received much attention, as it lends itself readily for combinations with vision and other modalities (Lederman and Klatzky, 2004). Botvinick and Cohen (1998) have shown that vision can change proprioceptive perception by way of touch. In their experiments, a group of subjects were asked to place their left arms on a table. This arm was shielded from the subjects' vision by a screen, but a plastic arm was placed in the visual field of the subject instead. While subjects were looking at the plastic arm, the left hand of the subject and the plastic arm were simultaneously touched with a brush. The subjects perceived that the plastic arm belonged to them, and the touch sensation came from the plastic arm itself (Botvinick and Cohen, 1998). Johnson, Burton and Ro (2006) used an apparatus connected to the hands of their subjects to give light and touch sensations at the same time. They have observed that sensations in one modality can amplify the sensations in the other modality, allowing weaker sensations to be perceived, but also making it possible to induce a sense of touch by just visual input.

The crossmodal interaction between touch and hearing was shown with an experiment by Jousmäki and Hari (1998), in what they have termed the *parchment illusion*. In this experiment, the subjects were asked to rub their hands, while listening to the produced sound from a headphone. The experiment demonstrated that by changing the high-frequency components of the sound it was possible to change the induced feeling of touch, making the subjects feel their hands to be dry or moisturized.

The demonstration of crossmodal influences between touch and smell go back a long time. In a paper published in 1932, Laird has shown that women assessing the quality of silk stockings by touch were influenced by the smell of the product. Among identical stockings, the one that smelled like flowers was consistently found to be of higher quality, but the subjects justified their decisions without referring to the sense of olfaction. Thus, this experiment showed an unconscious fusion of smell and touch. More recent research results along the same lines established that different smells can influence the perceived softness of different textiles (Demattè, Sanabria, Sugarman and Spence, 2006).

Fusion of visual and olfactory modalities was shown in an experiment that relied on naming smells with colours (Gilbert, Martin and Kemp, 1996). The experimenters found that certain smells were consistently associated with certain colours. Morrot, Brochet and Dubourdieu (2001) have coloured white wine with an odourless red colour, and asked wine experts to smell the wine. More than fifty professional wine tasters wrongly identified the wine as red wine. On the other hand, tasters who had their eyes closed during the experiment had 60-70 per cent success rate. This experiment demonstrates a visual influence on smell. Gottfried and Dolan (2003) have used fMRI experiments to investigate the relation between smell and sight, and found a facilitation of the smell perception in the presence of correlated visual input, possibly related to an increased activation in the anterior part of hippocampus and rostromedial orbitofrontal cortex.

In a series of experiments investigating the relation between gustatory and olfactory modalities Diamond and colleagues have established that some taste perceptions increase the sensitivity to smells with which they were associated (Diamond et al., 2005). For instance, the subjects who tasted saccharine became more sensitive to the smell of benzaldehyde. With a training that lasts a couple of weeks it was possible to teach the subjects new taste-smell combinations that would produce similar effects.

The majority of neurons in the taste- and smell-related areas in the orbitofrontal cortex code smell perceptions relative to the previously associated and learned taste sensations. Some of

these neurons respond to the smell, but also to the texture and density of the substance in the subjects mouth (Rolls, 2004). It has been shown that the sense of pleasure derived from eating regulates the fusion of smell, taste, texture and sight in these areas, and as the subject nears the point of satiation, the activation diminishes and finally disappears. Areas that are activated by both taste and smell cues are found in the caudal orbitofrontal cortex, amygdala, insular cortex and its adjoining areas, and anterior cingulate cortex (De Araujo et al., 2003).

NEUROLOGICAL FINDINGS

In crossmodal fusion, information from two or more perceptual modalities is processed, usually pre-consciously, to form a single perception. Multisensory neurons that are activated through multiple sensory channels play an important role in this process. Stein and Meredith (1993) have shown that the receptive fields of multisensory neurons were overlapping for different perceptual modalities.

Figure 2 symbolically shows the overlap between the auditory and visual receptive fields of two multisensory neurons in the superior colliculus (SC). Such topologically ordered areas are not only found in superior colliculus, but in many areas of the brain. Multisensory information is processed at the cortical level in superior temporal sulcus, intraparietal sulcus, parieto-preoccipital cortex, posterior insula, and the frontal cortex including premotor, prefrontal and anterior cingulate. At the sub-cortical level, claustrum, superior colliculus, supragenicular and medial pulvinar nuclei of the thalamus, rhinal cortex, amygdala-hippocampus show receptiveness to multisensory input (for a complete review of neurological findings, see Calvert, 2001). Among these, the superior colliculus is the most researched area in the context of multisensory fusion.

The primary role of SC is seen as bringing sensory modalities that are each represented in their own sense-specific coordinate systems into a single, common coordinate system. This system is responsible from coding the signals to be sent to the motor areas.

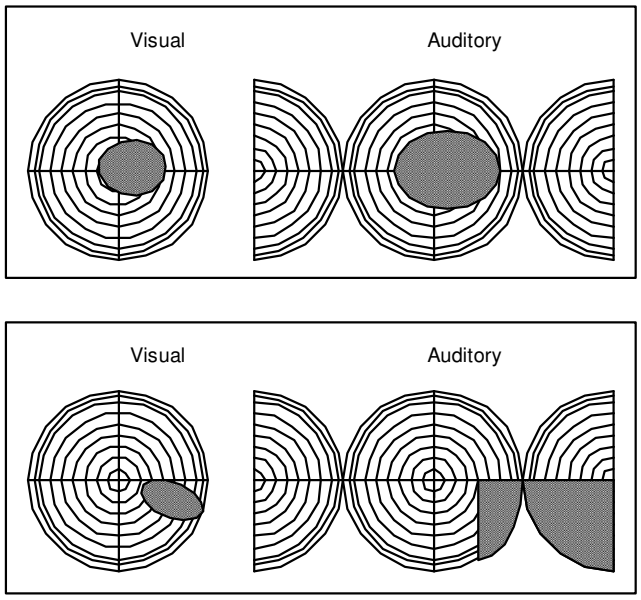

Figure 2 Receptive fields of two multisensory neurons in the superior colliculus (SC), shown in gray. Each concentric circle shows 10° of the receptive field, and the posterior part of the auditory field is shown by two half-circles (adapted from Stein et al. 2004).

According to the findings of Stein et al (2004), eighty per cent of the neurons in SC show activation levels close to the sum of activations from two modalities. If both modalities have signals coming from their receptive fields, the multisensory neuron shows activity that goes beyond the sum of the individual modalities. Thus, consistent multimodal signals are amplified. On the other hand, if one of the modalities has no activation in its receptive field, the activity is lower than the modality with the receptive-field activation. Hence, inconsistent multimodal signals are inhibited or suppressed.

One line of research investigates whether connections to SC already carry multimodal information. Anterior ectosylvian sulcus (AES) can be separated into three parts, related to visual, auditory and haptic information, respectively, and has strong projections to SC (Stein and Meredith, 1993). It has been shown that AEC also contains multisensory neurons similar to the ones in SC, but apart from these, connections to SC do not show multimodal responses (Wallace, Meredith and Stein, 1993).

To understand the relations between SC and AES, single-neuron activities from multisensory neurons in SC are recorded for audio, visual, and audio+visual input. Then AES is reversibly deactivation by cooling, and the measurements were repeated. Clemo and Stein (1986) found that responses in SC for audio and visual stimuli do not change when AES is de-activated, but

multisensory fusion of audio+visual stimuli was greatly impaired. Similar findings were also obtained with rostral suprasylvian sulcus (rLS).

An important property of fusion in multisensory neurons is that amplification of activity is greatest for cases where individual modalities have weak but consistent activation. For example, in a case where the visual input is so weak that it is not perceptible alone, it can be made perceptible by the presence of an auditory signal, even if the latter is also very weak. Single-neuron studies have confirmed behavioural studies of this phenomenon. We should note that this scheme is very different than information fusion approaches in computer systems, and it will be interesting to see whether it can be successfully applied to computer fusion.

PERCEPTUAL FUSION IN COMPUTER-BASED SYSTEMS

In this section, we look at how perceptual information is processed and fused in computer-based systems. Perceptual information in this context means digital information sampled from the environment via sensors of different sensitivity. These sensors can be cameras (vision), microphones (audio), haptic pressure sensors (touch), chemical sensors (smell and taste), thermometers, sonar-based proximity sensors, vibration sensors, passive infrared interruption sensors, etc. Even though the number, the range and the quality of sensors available to computers is very different than those available to humans, the approaches designed for information fusion can give normative insights for fusion performed by the brain.

In systems with multiple sensors, we can classify the relations between sensors into three groups (Durrant-Whyte, 1988):

- *Complementary sensors*: These sensors operate independently, and produce complementary information. The eye and the ear are complementary sensors.
- *Competitive sensors*: Sensors can have errors in their measurements. Competitive sensors measure the same phenomenon. Fusing the output of competitive sensors lowers the expected measurement error. Pressure sensors closely placed in a finger can be considered as competitive sensors.
- C*ooperative sensors*: Sometimes, a single sensor is not enough to measure a phenomenon, and the cooperation of several sensors is necessary. For instance depth perception is achieved with the cooperation of two eyes.

It is frequent practice in computer science to conceptualize processing elements as units with some input and some output. With some abuse of terminology, we can call a computational unit that takes in perceptual (sensory) information and produces a desired output (this can be a motor control signal, but also some semantic information) a *perceptual function.* To give an example, a perceptual function can receive visual and auditory input from a robot's sensory system, and generate the motor commands that would move the robot's arm to grasp an object.

The problem of finding the perceptual function that connects a particular perceptual input to a particular desired output is called (perceptual) *learning*. The outcome of learning is a model, often with a mathematical description, to associate a certain input with an output. The nature of the model is a major distinction in learning. In *parametric* models, the mathematical expression of the model is known, and the problem consists of finding the appropriate parameter values for the model. The nature of the model incorporates certain assumptions about the input, and its relation to the output. In *non-parametric* models, there are less assumptions about the nature (or distribution) of the input. Neural networks, frequently used in biologically-motivated models, are parametric models.

Learning systems used for perceptual problems are usually specialized for answering a single question. For instance in an artificial nose system, there are several questions one may ask: Is this gas poisonous? Is it flammable? Is it harmful? Is it of high quality? Does it contain a specific molecule? Usually for each question, a different learner needs to be constructed, but a learner can use the output of another learner as additional input.

In computer-based systems information fusion is necessary, primarily because information from a single sensor is not sufficient to answer a particular question and almost always contains noise. The most prevalent strategy is divide-and-conquer, where a problem is decomposed to its sub-problems, hopefully easier to solve than the entire problem. For instance in the artificial nose problem, different chemical sensors are used to address different aspects of the question, and the sensor readings are fused.

In learning systems, we can talk of three stages where information can be fused (Hall and Linas, 1997). These are *data-level*, *feature-level* and *decision-level* fusion. All these approaches can be used in biologically motivated models, either individually, or jointly.

In the data level fusion, data are fused as they come from the sensors, and passed jointly to the next stage of processing. In this approach, the sensor readings must be similar in range, or they must be normalized. In the human neural system, all sensory input is coded as neural signals, and normalization is not necessary. An example of data-level fusion in computer-based systems would be concatenating multiple pressure sensor inputs and treating them as a single, multi-dimensional sensor reading.

After the data acquisition phase, usually there is a pre-processing stage where data are prepared for the task of learning. Noise filtering, interpolation, smoothing are typical data pre-processing operations. But pre-processing goes further than simple filtering. A good transformation of data can greatly simplify the subsequent learning. We distinguish between *feature extraction* (transformations of data) and *feature selection* (selections of data dimensions) at this stage. To give an example from the human visual system, the simple cells in visual cortex V1 area are sensitive to oriented horizontal and vertical lines, whereas complex cells that receive input from many simple cells can have directional selectivity. The task they perform can be seen as feature extraction, in that their activation indicates the presence of particular features in the visual field. If the subject is looking for an object with horizontal features in his or her visual field, this sort of feature extraction will be useful. Simple perceptual features will be useful for many perceptual tasks, whereas complex features will help more specialized tasks.

Feature-level fusion relies on feature extraction from multiple sensory modalities. The joint feature vector can be used to train a learner. For some models, most notably for neural networks, the learning process becomes more time-consuming and difficult as the number of input features is increased. On the other hand, using more features with a single learner can potentially allow the learner to model the correlations between these features, leading to more accurate models.

Decision-level fusion happens when multiple learners individually give decisions for the output, and these outputs are combined to give a single output for the whole system. This procedure is especially useful if the learners consider different aspects of the input, and bring some *diversity* to the system. The gain from decision-level fusion is maximized if the errors of the learners are uncorrelated. In this approach, the individual decisions can be summed up, a majority vote can be conducted, the certainty of learners can be taken into account for a weighted vote, or even a meta-learner can be trained to make sense of the decisions of the individual learners. Figure 3 shows the basic fusion approaches schematically. Kuncheva

(2004) further splits decision-level fusion into two sub-branches, adding *classifier-level* fusion to the list. In this approach, different types of learners are used to produce decisions for the same input, so that the differences inherent in the learning approaches can contribute to the diversity.

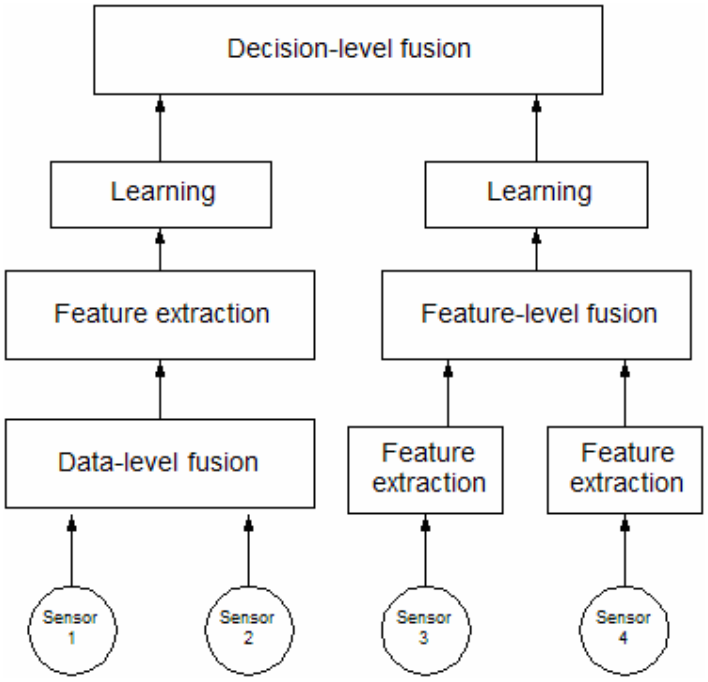

Figure 3 Principal fusion approaches

The *training* of a learner involves the use of a *training set*, a set of data on which the parameters of the learner are determined. The training set is decisive in the behaviour of the learner. Some fusion approaches use different training sets with the same learning model to produce a diverse set of learners. If the training sets are sufficiently diverse, the learners can produce very different outputs. To give an example from human visual system, consider the *other race effect* observed in human face recognition. Humans have difficulty in distinguishing individuals of a race with different and unfamiliar facial characteristics. The reason for this confusion is that the training set we get (i.e. faces we see as we grow up) contains the facial variations of our race, which are learned, but not the variations peculiar to a different race. Valentine (1991) has modeled this phenomenon in a computer-based face recognition system by using principal components analysis to learn the variations in a particular set of faces, and obtained an “other race effect” on a set with different dimensions of facial variation.

In information fusion, all these approaches can be used individually or jointly (Dasarathy, 1997). The perceptual input from sensors can be fed to sensor-specific learners, as well as to a general learner that operates on all sensory input, and then all the learners' outputs can be fused at the decision level. It is possible to use a learner to make the problem more manageable by eliminating several options at an early stage, and train subsequent learners on a simplified problem (Gökberk, Salah and Akarun, 2005). A very important difference between the brain and computer systems is that the brain computes in a massively parallel fashion, whereas the computer mostly computes serially. However, computers can simulate a parallel system at a given resolution. Therefore, it is possible to inspect the behaviour of a parallel system with the computer.

DISCUSSION

The active research areas in crossmodal fusion are how the brain achieves the integration of sensory input, which neural correlates are responsible from this process, and how the perceptual binding is achieved. Computer models can help in answering these questions, even though they are very simplified models. For instance *crosstalk* between two sensory modalities can even be observed in a simple neural network that receives multimodal input. The reason is that the neural network models the statistical relations in the input to determine the output, and doesn't assume anything about the source of the input. If two perceptual modalities are correlated, the neural network will make use of this connection automatically. For instance, taste and smell are strongly correlated modalities, and it is natural to expect strong crossmodal fusion between them, just for statistical reasons. Once computer models are employed to explain the bulk of the interactions, neurological studies can establish what other mechanisms are necessary for the whole range of behavioural observations.

Many hypotheses put forward to explain how the brain solves certain problems implicitly assume that the brain works optimally under certain conditions. For instance Bock (1986) claimed that the space occupied by a perceived object is coded with an eye-centered coordinate system in all perceptual modalities. Pouget, Deneve and Duhamel (2005) objected to this hypothesis on the grounds that this coding would not be optimal. The idea that the organization and processing of the brain should be optimal (or near-optimal) surfaces in many biologically-motivated models. Even though there are no theoretical reasons justifying this belief, successful systems and models were obtained by mimicking the brain in practice. On the other hand, computational arguments are frequently used to support a decision between different cognitive hypotheses to explain brain functionality.

Computer models for processing perceptual information may serve two different goals. The first goal is to solve a perceptual problem as well as humans. The yardstick to measure the success of these systems is the performance and accuracy of the system. The second goal is to solve a perceptual problem *the way humans solve it*. These systems work with information processing models that are abstractions of brain functionality at some level, and simulate the information flow to gain insight about the components of the model. The abstraction can be at a very simple signal-processing level (e.g. using Gabor wavelet filters that model receptive fields of simple cells in lateral geniculate nucleus for image processing; Daugman 1985), or at a higher level (e.g. simulation of the selective attention mechanism for visual pattern recognition; Salah, Alpaydın, and Akarun, 2002). The analysis of the computer models for descriptive purposes must take into account the fundamental differences between the brain and the computers, no matter what level of abstraction is selected. For instance perceptual fusion in the human brain results in faster response times, whereas in computers the response time is slower due to the increased processing load. Understanding the similarities and differences of information fusion in the brain and in computers is essential for employing computer models fruitfully.

CONCLUSION

In the discussions following the Biologically Inspired Information Fusion Workshop held in Surrey in 2006 (BIIF, 2006), the high-priority research areas on perceptual fusion were grouped under three headings:

1. Clinical applications of perceptual fusion: Inspecting influences of perceptual fusion on neurological and psychological disorders to develop prostheses and devices for the patients.
2. Applications of biological processes that achieve crossmodal fusion: Gaining insight for computer modeling of perceptual fusion. Understanding the benefits and the cost of fusing two modalities for different application areas.
3. Preparation of a common conceptual framework and a common terminology to allow interdisciplinary work on crossmodal fusion.

Theoretical and practical cross-pollination between disciplines, as well as opening of new application areas are targeted with research in these areas.

ACKNOWLEDGEMENTS

I thank Prof. Dr. Nil Molinas Mandel for her contributions.

REFERENCES

[1] Andersen, T.S., K. Tiippana, M. Sams, "Factors influencing audiovisual fission and fusion illusions," *Cognitive Brain Research*, Vol. 21(3), pp.301-308, 2004.

[2] De Araujo, I.E.T., E.T. Rolls, M.L. Kringelbach, F. McGlone, N. Phillips, "Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain," *European Journal of Neuroscience,* vol.18, pp.2059-2068, 2003.

[3] Bertelson, P., "Ventriloquism: a case of cross-modal perceptual grouping," in G. Aschersleben, T. Bachmann, J. Müsseler (eds.), *Cognitive contributions to the perception of spatial and temporal events,* Elsevier, 1999.

[4] Bertelson, P., B. de Gelder, "The psychology of multimodal perception," in C. Spence, J. Driver (eds.), *Crossmodal Space and Crossmodal Attention*, Oxford Univ. Press, 2004.

[5] BIIF - Biologically Inspired Information Fusion Workshop, Surrey, August 2006.

[6] Bischoff, M., B. Walter, C.R. Blecker, K. Morgen, D. Vaitl, G. Sammer, "Utilizing the ventriloquism-effect to investigate audio-visual binding," *Neuropsychologia*, yayında, 2006.

[7] Bock, O., "Contribution of retinal versus extraretinal signals towards visual localization of goal directed movements," *Experimental Brain Research*, vol.64, pp.476-482, 1986.

[8] Botvinick, M., J. Cohen, "Rubber hands 'feel' touch that eyes see," *Nature*, vol.391, pp.756, 1998.

[9] Calvert, G.A., "Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies," *Cerebral Cortex*, vol.11(12), pp.1110-1123, 2001.

[10] Calvert, G.A., E. Bullmore, M. Brammer, R. Campbell, S. Williams, P. McGuire, P., Woodruff, S. Iversen, A. David, "Activation of auditory cortex during silent lipreading," *Science*, vol.276, pp.593-596, 1997.

[11] Clemo, H.R., B.E. Stein, "Effects of cooling somatosensory cortex on response properties of tactile cells in the superior colliculus," *Journal of Neurophysiology,* vol.55, pp.1352-1368, 1986.

[12] Dasarathy, B., "Sensor Fusion Potential Exploitation - Innovative Architectures and Illustrative Applications," *Proceedings of the IEEE*, vol.85(1), pp.24-38, 1997.

[13] Daugman, J.G., "Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol. 2, pp.1160-1169, 1985.

[14] Diamond, J., P.A.S. Breslin, N. Doolittle, H. Nagata, P. Dalton, "Flavor Processing: Perceptual and Cognitive Factors in Multi-modal Integration," *Chemical Senses*, vol.30 (suppl.1), pp.232–233, 2005.

[15] Driver, J., "Enhancement of listening by illusory mislocation of speech sounds due to lip-reading," *Nature,* vol.381, pp.66-68, 1996.

[16] Driver, J., Spence, C., "Attention and the crossmodal construction of space," *Trends in Cognitive Sciences*, vol.2, pp.254-262, 1998.

[17] Durrant-Whyte, H.F., "Sensor models and multisensor integration," *International Journal of Robotics Research,* vol.7(6), pp.97-113, 1988.

[18] Fowler, C.A., D.J. Dekle, "Listening with eye and hand: cross-modal contributions to speech perception," *Journal of experimental psychology: Human perception and performance*, vol.17(3), pp. 816-828, 1991.

[19] Giard, M.-H., F. Peronnet, "Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study," *Journal of Cognitive Neuroscience*, vol.11, pp.473–490, 1999.

[20] Gilbert, A.N., R. Martin, S.E. Kemp, "Cross-modal correspondence between vision and olfaction: the color of smells," *American Journal of Psychology*, vol.109, pp.335–351, 1996.

[21] Gottfried, J.A., R.J. Dolan, "The Nose Smells What the Eye Sees: Crossmodal Visual Facilitation of Human Olfactory Perception," *Neuron*, vol.39(2), pp.375-386, 2003.

[22] Gökberk, B., A.A. Salah, L. Akarun, "Rank-based Decision Fusion for 3D Shape-based Face Recognition," in T. Kanade, A. Jain, N. Ratha (eds.) Lecture Notes in Computer Science, Volume 3546/2005, *AVBPA*, pp.1019-1028, 2005.

[23] Hall, D.L., J. Linas, "An introduction to multisensor data fusion," *Proceedings of the IEEE,* vol.85(1), pp.6-23, 1997.

[24] Hay, J.C., H.L. Pick, K. Ikeda, "Visual capture produced by prism spectacles," *Psychonomic Science,* vol.2, pp.215-216, 1965.

[25] Johnson, R.M., P.C. Burton, T. Ro, "Visually induced feelings of touch," *Brain Research*, vol.1073-1074, pp.398-406, 2006.

[26] Jousmäki V, R. Hari, "Parchment-skin illusion: sound-biased touch," *Current Biology*, vol.8(6), pp.190, 1998.

[27] Knudsen, E.I., M.S. Brainard, "Visual instruction of the neural map of auditory space in the developing optic tectum," *Science*, vol.253, pp.85-87, 1991.

[28] Knudsen, E.I., P.F. Knudsen, "Vision guides the adjustment of auditory localization in young barn owls," *Science*, vol.230, pp.545-548, 1985.

[29] Kuncheva, L.I., *Combining Pattern Classifiers: Methods and Algorithms,* John Wiley & Sons, New Jersey, 2004.

[30] Laird, D.A., "How the consumer estimates quality by subconscious sensory impressions: with special reference to the role of smell", *Journal of Applied Psychology*, vol.16, pp.241–246, 1932.

[31] Lederman, S.J., R.L. Klatzky, "Multisensory Texture Perception," in G. Calvert, C. Spence, and B. Stein (eds). *Handbook of Multisensory Processes*. Cambridge: MIT Press, 2004.

[32] Massaro, D.W., M. Tsuzaki, M. Cohen, A. Gesi, R. Heridia, "Bimodal speech perception: An examination across languages," *Journal of Phonetics*, vol.21, pp.445-478, 1993.

[33] McGurk, H., J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol.407, pp.906-908, 1976.

[34] Morrot, G., F. Brochet, D. Dubourdieu, "The color of odors," *Brain and Language*, vol.79, pp.309–320, 2001.

[35] Murphy, R.R., "Biological and cognitive foundations of intelligent sensor fusion," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol.26(1), pp.42-51, 1996.

[36] Pouget, A., S. Deneve, J.-R. Duhamel, "A computational neural theory of multisensory spatial representations," in C. Spence, J. Driver (eds.), *Crossmodal Space and Crossmodal Attention*, Oxford Univ. Press, 2004.

[37] Radeau, M., P. Bertelson, "The effect of a textured visual field on modality dominance in a ventriloquism simulation," *Perception and Psychophysics,* vol.20, pp.227-235, 1976.

[38] Rock, I., C.S. Harris, "Vision and touch," *Scientific American*, vol.248, pp.96-107, 1967.

[39] Rolls, E.T., "Smell, taste, texture and temperature multimodal representations in the brain, and their relevance to the control of appetite," *Nutrition Reviews,* vol.62, pp.193-204, 2004.

[40] Rosenblum, L.D., M.A. Schmuckler, J.A. Johnson, "The McGurk effect in infants", *Perception & Psychophysics*, vol.59(3), pp. 347-357, 1997.

[41] Salah, A.A., E. Alpaydın, L. Akarun, "A Selective Attention Based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.24(3), pp.420-425, 2002.

[42] Schwartz, J.-L., J. Robert-Ribes, P. Escudier, "Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception," R. Campbell (Ed.), *Hearing by Eye: The Psychology of Lipreading*, Lawrence Erlbaum Associates, Hove, UK, pp. 3–51, 1998.

[43] Sekuler, R., A.B. Sekuler, R. Lau, "Sound alters visual motion perception," *Nature,* vol.385, p.308, 1997.

[44] Shams, L., Y. Kamitani, S. Shimojo, "What you see is what you hear," *Nature,* vol.408, p.788, 2000.

[45] Shams, L., Y. Kamitani, S. Shimojo, "Visual illusion iduced by sound," *Cognitive Brain Research,* vol.14, p.147-152, 2002.

[46] Stein, B.E., M.A. Meredith, *The merging of the senses*, Cambridge: MIT Press, 1993.

[47] Stein, B.E., T.R. Stanford, M.T. Wallace, J.W. Vaughan, W. Jiang, "Crossmodal spatial interactions in subcortical and cortical circuits," in C. Spence, J. Driver (eds.), *Crossmodal Space and Crossmodal Attention*, Oxford Univ. Press, 2004.

[48] Talsma, D., M.G. Woldorff, "Selective Attention and Multisensory Integration: Multiple Phases of Effects on the Evoked Brain Activity," *Journal of Cognitive Neuroscience*, vol.17(7), pp.1098-1114, 2005.

[49] Valentine, T., "A unified account of the effects of distinctiveness, inversion and race in face recognition," *Quarterly Journal of Experimental Psychology*, vol.43A, pp.161-204, 1991.

[50] Wallace, M.T., M.A. Meredith, B.E. Stein, "Converging influences from visual, auditory, and somatosensory cortices onto output neurons of the superior colliculus," *Journal of Neurophysiology*, vol.69, pp.1797-1809, 1993.

[51] Watanabe K, S. Shimojo, "Attentional modulation in perception of visual motion events," *Perception*, vol.27(9), pp.1041-54, 1998.

[52] Welch, R.B., D.H. Warren, "Immediate perceptual response to intersensory discrepancy," *Psychological Bulletin,* vol.88, pp. 638–667, 1980.