as in (Reeke, Sporns, & Edelman, 1990; Edelman, Reeke, & Gall, 1992), however, we concentrate solely on classification, but attack real classification tasks with overlapping input patterns. The fact that both networks are learning makes this approach significantly harder than approaches where one modality trains another (Munro, 1988; Carpenter, Grossberg, & Reynolds, 1991; Tan 1995) or others that combine two already trained networks (Yuhas, Goldstein, & Sejnowski, 1988; Stork, Wolff, & Levine, 1992).

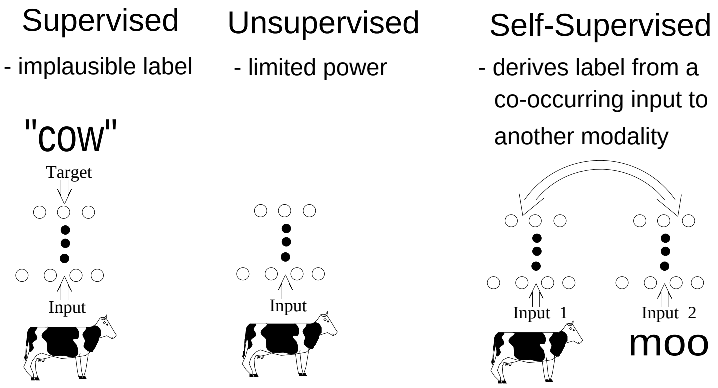

Figure 3: **The idea behind the self-supervised algorithm.**

To start we give some of the biological motivation behind the idea of using the cross-modality structure.

2.1 Psychophysics

The coincidence between auditory and visual events is pervasive throughout our experience. From just after birth, babies are able to turn crudely toward sounds. Over time we develop a rather precise auditory-visual spatial map. This enables us to be able to pick up correlations between auditory and visual sensations. For example, with sound localization we are able to notice that mooing comes from cows.

Another example of information picked up in this way is the ability to read lips. Anyone who has conversed in a noisy environment, is aware of the improved speech recognition achieved when the speaker's face (particularly the lips) is visible. (This has also been demonstrated in more controlled experimental conditions (Sumby, & Pollack, 1954).) The visual signal from the motion of the lips, jaw, and tongue help the auditory system to understand the speech.

This ability to recognize relationships between lip movements and emitted sound develops early. By four and a half months of age infants are able to recognize that particular lip motions go with particular sounds. Kuhl and Meltzoff (1984) showed that infants looked significantly longer at the matching face when presented with the sound /a/ or /i/. Their preference was specific to the actual speech information as they did not show this effect when the speech signals were replaced with tones that followed the duration, amplitude envelope and onset/offset timing of the original speech sounds (Kuhl & Meltzoff, 1984).

The effect of lip movement on speech recognition is even more prominent when the stimuli are experimentally manipulated so that the visual and acoustic signals are discordant. In the experiments of (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978) subjects are presented with acoustic stimuli of various consonant-vowel pairs and simultaneously shown images of faces speaking a different consonant with the vowel. Thus for example when presented acoustically with a /ba/ syllable and visually with a face speaking /ga/, 98% of adult subjects hear /da/ (McGurk & MacDonald, 1976). The result is very striking and not subject to conscious control. It shows that visual and auditory stimuli are able to interact to produce a unified percept, different from the stimuli actually given to either modality. Furthermore it seems that the ability of visual signals to influence acoustic classification is at least partially learned. Pre-school and school children show significantly less of the effect than do adults (McGurk & MacDonald, 1976).

Another example of experienced auditory-visual correlations affecting an auditory perception is given as an anecdote in Howells (1944). Howells reports that a driver who was familiar with an intersection that made a concurrent whistle sound as the lights changed for stop and go, failed to notice when the whistle was later disabled.

> On the next trip to the crossing, in spite of counter suggestions recently received, the driver reported a distinct hallucination of the whistle (Howells, 1944, p.89).

In another anecdotal example where the frequencies of the whistles were different for lights changing to the stop or go signal. Howells reports

> Few drivers noticed that there was a difference in the tones, or even that a whistle sounded at all, until the wires controlling the whistles were accidentally crossed by a repairman, so the usual combinations of color and tone were reversed. The result was general confusion and a collision at the crossing (Howells, 1944 p. 90).

Under more controlled conditions, Howells (1944) trained subjects with tones followed by colored screens. On 95% of the trials, the color-tone pairing was consistent and on the other 5% of trials it was reversed. After an initial period, half the presentations were at full saturation and the other half much paler. Subjects showed increasing errors (on the pale screens preceded by the inconsistent tone) with training. Subjects tested with white stimuli after conditioning reported always the color associated with the co-occurring sound and subjects instructed to produce a white color in the presence of the high or low tone were offset in the direction opposite to that imposed during training.

A similar example of auditory events influencing visual perception is demonstrated in a cross-modal experiment in (Durgin, 1995; Durgin & Proffitt, 1996). The experiment involved repeated brief presentations of random dot patterns in two rectangular areas of a screen. On each presentation, one of the two areas received 25 dots/deg^2 and the other 2 dots/deg^2. The visual presentations were paired with auditory tone stimuli such that the pitch of the tone was perfectly correlated with the side of the denser dot pattern. After 180 flashed presentations, a staircase procedure was used to determine the perceived density equivalence (for test patterns with dot densities between the two trained densities) between the two areas when presented with each of the two tones. The experiment showed that there was a significant effect of the tone on the perceived density relationship between the patterns in the two areas. The simultaneous presentation of the tone associated with a denser texture in one area during training, lead to an impression of greater dot density in that area during testing. To match a constant density, the difference between the density required in the presence of the high pitch and that with a low pitch was 10% (Durgin, 1995).

Hefferline and Perera (1963) have shown that correlated proprioceptive (an invisibly small thumb twitch detected electromyographically) and auditory events (tone) can lead to a subject later reporting that he “still heard it (the tone)” to subsequent proprioceptive events in the absence of the tone.

Zellner and Kautz (1990) have also shown that color can affect odor perception. In their experiments, colored solutions were perceived as having a more intense odor.

> Even after being told the solutions were of equal concentrations, they [subjects] insisted that the solutions were not the same intensity (Zellner and Kautz, 1990, p. 396)

It is clear that the co-occurrence of multi-sensory signals can assist or interfere with processing. There is also evidence that after experience with cross-modal correlations, a uni-modal discrimination can be affected by a stimulus to the other modality. We would like to go one step farther and hypothesize that this multi-modality integration is important not only for improved recognition but is useful for the development of recognition features in both individual modalities. Along these lines, there is some evidence that exposure to auditory-visual co-occurrences is important for normal attentional development. Quittner and colleagues (Quittner, Smith, Osberger, & Mitchell, 1994) report that deaf children show reduced *visual* attention (in a non-auditory task) than hearing children. The authors conclude that though auditory information is not used in their tested task the development of focused visual attention is helped by auditory experience (presumably coincident with visual experience).

2.2 Neurobiology

The previous section examined results showing that information from different sensory modalities is combined in determining our perception. Often, the combination is not subject to conscious control. It is as if the results are not simply being combined at a high-level output stage but are able to influence each other in the individual processing stages. This is corroborated by neurophysiological studies which have found responses of cortical cells in primary sensory areas that respond to features from other sensory modalities. For example Spinelli, Starr and Barrett (1968) found sound frequency specificity in cells in primary *visual* cortex of the cat and Fishman and Michael (1973) found that these bimodal cells tend to be clustered together. As support for the unified percept observed in psychophysical studies, the stimuli are able to affect the same cell. In fact acoustic responses in a single cell could be inhibited by inhibitory visual stimuli (Morrell, 1972). More recently Maunsell and colleagues (Haenny, Maunsell, & Schiller, **?**; Maunsell, Sclar, Nealey, & DePriest,1991) have shown responses in visual neurons in Area V4 to oriented tactile stimuli that the animal has been trained to match to subsequently presented oriented visual gratings.

Sams and colleagues, (Sams, Aulanko, Hämääinen, Hari, Lounasmaa, Lu,& Simola, 1991) have also shown effects of visual input on auditory processing in humans. Using magnetoencephalographic (MEG) recordings, they showed that although a visual signal by itself did not result in a response over the auditory cortical area, different visual signals changed the response to the auditory signal. They again used the McGurk effect stimuli. Subjects were trained with a higher percentage of either agreeing or disagreeing stimuli. Significantly different neuromagnetic measurements were made to the frequent and infrequent stimuli. As no similar difference occurred when two different light stimuli occurred with the sounds, they argue that this shows that the visual information from the face is reaching the auditory cortex.

The neurophysiological and psychophysical evidence must be reconciled with the fact that anatomically the information from the different sensory modalities goes to spatially separate, segregated cortical areas. Retinal input goes to occipital cortex whereas auditory input goes to auditory cortex in the temporal lobe. Even within the auditory and visual cortex there are many different areas which seem to be specialized to processing different parts of the signal. For instance color processing seems to be mostly separate from motion processing (Merigan & Maunsell, 1993; Desimone & Ungerleider, 1989). There is a significant restriction on the amount of cross-modality interaction that can occur. This is thought to be due to restrictions on connectivity; it is not physically possible to have all neurons connected to all other neurons (or even any significant fraction). Therefore input from each modality and submodality must be processed separately, at least in the early stages, with little cross-talk.

As there are no direct afferent (feed-forward) connections from one input modality to another, the information from other modalities could either be coming bottom-up from shared subcortical structures such as the superior colliculus or alternatively top-down from the multi-sensory integration areas such as entorhinal cortex and other limbic polymodal areas. This idea has been suggested before (for example Rolls, 1989) and seems to be supported by the evidence from visual cortex. As stated by Spinelli et al. (1968)

> non-visual stimuli affect the activity of ganglion cells only minutely (Spinelli, Pribram, & Weingarten, 1965; Spinelli & Weingarten, 1966; Spinelli, 1967; they affect that of the geniculate cells to a greater extent (Meulders, Colle, & Biosacq-Schepens, 1965) and very markedly affect cortical cells (Murata, Cramer, & Bach-y-Rita, 1965). Even more interaction appears to be present in prestriate cortex (Buser & Borenstein, 1959).(p. 82)

Thus we know from psychophysical studies that information from different modalities is combined and that information from one modality can assist or interfere with classification in another. The physiological evidence supports this finding in showing that input to other modalities can influence processing in another sensory pathway. This combined with the anatomical evidence that shows no direct input from one modality's transducers to another pathway, suggests that this information is coming top-down through feedback pathways from multi-sensory areas. Furthermore we suggest that this integration may not just affect the properties of developed systems but play an important role in the learning process itself. Just as lip-reading is a learned classification ability, correlations between inputs to different sensory modalities may affect other classification learning in the individual modalities. In this section we will investigate the power that this kind of integration might provide to learning classifiers in the individual modalities.

2.3 Using Cross-Modality Information for Self-Supervised Learning

Following the anatomical evidence presented earlier and acting under the assumption that it is infeasible to have neurons receiving input from the sensory transducers of all the senses, we propose an architecture such as that schematized in Figure 3 in which each modality has its own processing stream (or classification network) but access to each other's output at a high level. This information can reach the lower levels within each processing stream through feedback within the stream.

One way to make use of the cross-modality structure in a network like this is to cluster the codebook vectors in their individual spaces, but use the joint structure to learn to label co-occurring codebook vectors with the same label. After clustering in the input space, the activity patterns in the resulting hidden unit (codebook) activation space (the space of dimensionality equal to the sum of the number of codebook vectors in each space) can be clustered. For example using Competitive learning on the second layer of weights in a network such as that in Figure 4, will tend to cluster the codebook vectors. Each codebook vector is given the label of the output neuron in whose cluster it belongs (the neuron to which it projects most strongly); thus these weights can be considered implicit labeling weights.

The above use of cross-modality structure is useful and we will use it to initialize our algorithm (calling it the initial labeling algorithm), however a more powerful use of the extra information is for better placement of the codebook vectors themselves. The insight in this algorithm is that we

References

Artola, A., Bröcher, S., & Singer, W. (1990, September). Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, *347*, 69–72.

Becker, S. (1993). Learning to categorize objects using temporal coherence. In C. Giles, S.J.Hanson, & J. Cowan (Eds.), *Advances in Neural Information Processing Systems 5* (pp. 361—368). Morgan Kaufmann.

Becker, S., & Hinton, G. E. (1992, January). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*, 161–163.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Buser, P., & Borenstein, P. (1959). Responses somesthesiques, visuel et auditives, recuellies, au niveau du cortex "associatif" infrasylvien chez le chat curarise non anesthesie. *Electroencephalog. Clin. Neurophysiol.*, *11*, 285–304.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991). Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, *4*, 565–588.

Coultrip, R., Granger, R., & Lynch, G. (1992). A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks*, *5*, 47–54.

Desimone, R., & Ungerleider, L. G. (1989). Neural mechanisms of visual processing in monkeysIn F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology* (Vol. 2). Elsevier Science Publishers B. V.

Diamantini, C., & Spalvieri, A. (1995, November). Pattern classification by the bayes machine. *Electronics Letters*, *31*(24), 2086–2088.

Durgin, F. H. (1995). *Contingent aftereffects of texture density: Perceptual learning and contingency.* Unpublished doctoral dissertation, Department of Psychology, University of Virginia.

Durgin, F. H., & Proffitt, D. R. (1996). Visual learning in the perception of texture: Simple and contingent aftereffects of texture density. *Spatial Vision*, *9*(4), 423–474.

Edelman, G. M., Jr., G. N. R., Gall, W. E., Tononi, G., & Williams, D. (1992, August). Synthetic neural modeling applied to a real-world artifact. *Proc. Natl. Acad. Sci.*, *89*, 7267–7271.

Fishman, M. C., & Michael, C. R. (1973). Integration of auditory information in the cat's visual cortex. *Vision Research*, *13*, 1415–1419.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200.

George N. Reeke, J., Sporns, O., & Edelman, G. M. (1990, September). Synthetic neural modeling: The "darwin" series of recognition automata. *Proceedings of the IEEE*, *78*, 1498–1530.

Goldstone, R., & Schyns, P. (1994). Learning new features of representation. In *Proceedings of the 16th Annual Meeting of the Cognitive Science Society* (pp. 974–978). Erlbaum, Hillsdale NJ.

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*, 187–202.

Hebb, D. O. (1949). *The Organization of Behavior.* Wiley.

Hefferline, R. F., & Perera, T. B. (1963, march). Proprioceptive discrimination of a covert operant without its observation by the subject. *Science*, *139*, 834–835.

Howells, T. (1944). The experimental development of color-tone synesthesia. *Journal of Experimental Psychology*, *34*(2), 87—103.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Kohonen, T. (1990). Improved versions of learning vector quantization. In *IJCNN International Joint Conference on Neural Networks* (Vol. 1, pp. I–545–I–550).

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*, 361—381.

Lehky, S., & Sejnowski, T. (1988). Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature*, *333*, 452–454.

Linsker, R. (1986a, November). From basic network principles to neural architecture: Emergence of orientation-selective cells. *Proc. Natl. Acad. Sci. USA*, *83*, 8390–8394.

Linsker, R. (1986b, November). From basic network principles to neural architecture: Emergence of orientation columns. *Proc. Natl. Acad. Sci. USA*, *83*, 8779–8783.

Linsker, R. (1986c, October). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proc. Natl. Acad. Sci. USA*, *83*, 7508–7512.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–257.

Maunsell, J., Sclar, G., Nealey, T., & DePriest, D. (1991). Extraretinal representations in area V4 of macaque monkey. *Visual Neuroscience*, *7*(6), 561–573.

McGurk, H., & MacDonald, J. (1976, December). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Merigan, W. H., & Maunsell, J. H. R. (1993). How parallel are the primate visual pathways? In *Annual Review of Neuroscience* (Vol. 16, pp. 369–402).

Meulders, M., Colle, J., & Biosacq-Schepens, N. (1965). Macro and microelectrode studies of somatic responses in the lateral geniculate body. In *Proceedings, XXIII International Congress of Physiological Sciences* (p. 364).

Miikkulainen, R. (1991). Self-organizing process based on lateral inhibition and synaptic resource redistribution. In T. Kohonen, K. Makisära, O. Simula, & J. Kangas (Eds.), *Artificial Neural Networks* (pp. 415–420). Elsevier Science Publishers.

Miller, G. A., & Nicely, P. E. (1955, March). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989, August). Ocular dominance column development: Analysis and simulation. *Science*, *245*, 605–615.

Morrell, F. (1972, July). Visual system's view of acoustic space. *Nature*, *238*, 44–46.

Munro, P. (1988, January). *Self-supervised Learning of Concepts by Single Units and "Weakly Local" Representations* (Report No. LIS003/IS88003).

Murata, K., Cramer, H., & Bach-y-Rita, P. (1965). Neuronal convergence of noxious, acoustic and visual stimuli in the visual cortex of the cat. *Journal of Neurophysiology*, *28*, 1233–1239.

Nowlan, S. J. (1991). *Soft Competitive Adaptation: Neural Network Learning Algorithms based on Fitting Statistical Mixtures.* Unpublished doctoral dissertation, School of Computer Science, Carnegie Mellon University.

Obermayer, K., Ritter, H., & Schulten, K. (1990, November). A principle for the formation of the spatial structure of cortical feature maps. *Proc. Natl. Acad. Sci.*, *87*, 8345–8349.

Obermayer, K., Schulten, K., & Blasdel, G. (1992). A comparison between a neural network model for the formation of brain maps and experimental data. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in Neural Information Processing Systems 4* (pp. 83–90). Morgan Kaufmann.

Peterson, G., & Barney, H. (1952). Control methods used in a study of vowels. *The Journal of the Acoustical Society of America*, *24*, 175–184.

Phillips, W., Kay, J., & Smyth, D. (1995). The discovery of structure by multi-stream networks of local processors with contextual guidance. *Network: Computation in Neural Systems*, *6*, 225—246.

Polana, R. (1994). *Temporal Texture and Activity Recognition.* Unpublished doctoral dissertation, Department of Computer Science, University of Rochester.

Quittner, A., Smith, L., Osberger, M., Mitchell, T., & Katz, D. (1994). The impact of audition on the development of visual attention. *Psychological Science*, *5*(6), 347—353.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, *5*, 289–304.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Math. Stat.*, *22*, 400—407.

Rolls, E. (1989). The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampusIn R. Durbin, C. Miall, & G. Mitchison (Eds.), *The Computing Neuron* (). Addison-Wesley.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533—536.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1, pp. 151–193). MIT Press.

Sa, V. de, & Ballard, D. (1992). Top-down teaching enables task-relevant classification with competitive learning. In *IJCNN International Joint Conference on Neural Networks* (Vol. 3, pp. III–364—III–371).

Sa, V. R. de. (1994a). Minimizing disagreement for self-supervised classification. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman, & A. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School* (pp. 300—307). Erlbaum Associates.

Sa, V. R. de. (1994b). *Unsupervised Classification Learning from Cross-Modal Environmental Structure.* Unpublished doctoral dissertation, Department of Computer Science, University of Rochester. also available as TR 536 (November 1994)

Sams, M., Aulanko, R., Hämääinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*, 141–145.

Sklansky, J., & Wassel, G. N. (1981). *Pattern Classifiers and Trainable Machines.* Springer-Verlag.

Spinelli, D. (1967). Receptive field organization of ganglion cells in the cat's retina. *Experimental Neurology*, *19*, 291–315.

Spinelli, D., Pribram, K., & Weingarten, M. (1965). Centrifugal optic nerve responses evoked by auditory and somatic stimulation. *Experimental Neurology*, *12*, 303–319.

Spinelli, D., Starr, A., & Barrett, T. W. (1968). Auditory specificity in unit recordings from cat's visual cortex. *Experimental Neurology*, *22*, 75–84.

Spinelli, D., & Weingarten, M. (1966). Afferent and efferent activity in single units of the cat's optic nerve. *Experimental Neurology*, *15*, 347–362.

Stork, D. G., Wolff, G., & Levine, E. (1992). Neural network lipreading system for improved speech recognition. In *IJCNN International Joint Conference on Neural Networks* (Vol. 2, pp. II–286—II–295).

Sumby, W., & Pollack, I. (1954, March). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, *26*(2), 212–215.

Tan, A.-H. (1995). Adaptive resonance associative map. *Neural Networks*, *8*(3), 437–446.

Wassel, G. N., & Sklansky, J. (1972). Training a one-dimensional classifier to minimize the probability of error. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(4), 533—541.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* Unpublished doctoral dissertation, Harvard University.

Yuhas, B., Jr., M. G., & Sejnowski, T. (1988). Neural network models of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, *78*(10), 1658—1668.

Zellner, D. A., & Kautz, M. A. (1990). Color affects perceived odor intensity. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 391–397.

Zipser, D., & Andersen, R. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*, 679–684.