# Biologically Motivated Audio-Visual Cue Integration for Object Categorization

(*) J. Anemüller, J-H. Bach, B. Caputo, L. Jie, F. Ohl, F. Orabona, R. Vogels, D. Weinshall, A. Zweig

***Abstract*— Auditory and visual cues are important sensor inputs for biological and artificial systems. They provide crucial information for navigating environments, recognizing categories, animals and people. How to combine effectively these two sensory channels is still an open issue. As a step towards this goal, this paper presents a comparison between three different multi-modal integration strategies, for audio-visual object category detection. We consider a high-level and a low-level cue integration approach, both biologically motivated, and we compare them with a mid-level cue integration scheme. All the three integration methods are based on the least square support vector machine algorithm, and state of the art audio and visual feature representations. We conducted experiments on two audio-visual object categories, dogs and guitars, presenting different visual and auditory characteristics. Results show that the high-level integration scheme consistently performs better than single cue methods, and of the other two integration schemes. These findings confirm results from the neuroscience. This suggests that the high-level integration scheme is the most suitable approach for multi-modal cue integration for artificial cognitive systems.**

## I. Introduction

Cognitive systems are intrinsically multi-modal. This is true for biological systems as well as for artificial ones. Multi-modality guarantees independent, diverse and information-rich sensory inputs, that make it possible robust performance in varied, unconstrained settings. An important, open issue is how to combine cues from diverse sensors, so to achieve optimal performance. This topic has been vastly investigated in the pattern recognition literature and in the field of the neurosciences (we refer the reader to section II for a review of the relevant literature in the field). From the algorithmic point of view, we can identify three main multi-modal integration strategies [1]: *(a) low-level integration*, where cues are combined at the feature level; *(b) mid-level integration*, where cues are combined together while building the classification decision function, and *(c) high-level integration*, where cues are used separately to produce confidence estimates, which are then combined together. Biological studies seem to indicate that integration happens at the highest level [2], [3], even if some results indicate that some form of integration happens also at the low-level [4].

In this paper we present a comparative evaluation on cue integration methods for audio-visual object category detection. We take a discriminative approach, and use a Least Square-Support Vector Machine [5] as the main building block for three different cue integration strategies. These strategies are respectively a low-level, a mid-level and a high-level integration scheme. Results show that the high level integration scheme outperforms the other two. This result is in agreement with a consistent body of literature in the neurosciences (see for instance [2] and references therein).

The rest of the paper is organized as follows: section II gives an overview of the state of the art in multi-modal cue integration for biological systems, and for pattern recognition algorithms. Section III describes the algorithms used for cue integration, and section IV illustrates the experimental setup (section IV-A), the audio-visual features used (section IV-B-IV-C), and the obtained results (section IV-D). The paper concludes with an overall discussion and possible directions for future research.

## II. Previous Work

There is plenty of evidence that integrating inputs from different sensory modalities can greatly enhance the ability of animals and humans to cope economically and flexibly with complex and ever-changing environments [6]. Despite its fundamental relevance, the neuroanatomical and neurophysiological bases for cross-modal integration are still not well understood. Within the neurosciences, three lines of research have emerged that tackle mechanisms and functional roles of multi-sensory integration. First, it has been known for a long time, from anatomical and physiological studies, that certain subcortical brain structures, like the superior colliculus, harbor neurons which are multi-modally sensitive, i.e. that can driven by stimuli from different modalities (e.g. visual and auditory) (for an overview on the topic see [2]). A second line of research, currently experiencing a revival, has demonstrated that even in sensory cortices that were traditionally considered as uni-sensory, integration of input from different sensory modalities does exist (see for instance [4], [7], [8]). While several anatomical candidate systems exist that are hypothesized to mediate these different forms of cross-modal integration, it is currently not clear which particular anatomical system supports which particular neurophysiological mechanism. However, there is some evidence for separate functional roles of these types. For example, in the superior colliculus only a relatively small

(*) In alphabetical order

This work was supported by the EU project DIRAC (FP6-0027787).

J. Anemüller, J.-H. Bach are with the University of Oldenburg, Germany {joerg-hendrik.bach, joern.anemueller}@uni-oldenburg.de

B. Caputo, L. Jie, F. Orabona are with the IDIAP Research Institute, Martigny, Switzerland {bcaputo,jluo,forabona}@idiap.ch

F. Ohl is with the University of Magdeburg, Germany frank.ohl@ifn-magdeburg.de

R. Vogel is with the Catholic University of Leuven, Belgium Rufin.Vogels@med.kuleuven.be

D. Weinshall, A. Zweig are with the Hebrew University of Jerusalem, Israel {daphna,alon.zweig}@cs.huji.ac.il

fraction of neurons shows cross-modal sensitivities, but these neurons can definitely be driven out of their resting state by stimuli from different modalities. In contrast, in sensory cortices multi-sensory sensitivity manifests itself more often in rather subtle modulations of activity, but this is found in a large fraction of neurons (see [9] for an overview). A relatively new line of research investigates the integration of stimuli from different sensory modalities not because they are presented in spatial and/or temporal proximity, but because of convergent meaning. For example, it has been demonstrated that cortical mechanisms exist that allow the transfer of knowledge about stimuli obtained in one sensory modality to novel, previously unexperienced stimuli from another sensory modality ([10]).

Many cue integration methods have been presented so far in the pattern recognition literature. For instance, Clark and Yuille [11] classify these methods into two main groups, *weak coupling* and *strong coupling*. Assuming that each cue is used as the input to a different classifier, weak coupling is when the output of two or more independent classifiers are combined. On the other hand, strong coupling is when the output of one classifier is affected by the output of another classifier, so that their output are no longer independent. In this paper we will focus on weak coupling. For this family of approaches, computing confidence estimates is a key issue. This is an open problem for discriminative classifiers. Although classifiers like K-NN, ANN, or SVM output numeric scores for class membership, some experiments show that, when used directly, they are not well correlated with classification confidence [12]. Several authors attacked this problem by developing more sophisticated measures such as probability estimates obtained by trained sigmoid function [13] with extensions for multi-class problems [14], or relative distance from the separating hyperplane, normalized with the average class distance from the plane [15].

Cue integration via accumulation was first proposed in a probabilistic framework by Poggio *et al.*[16], and then further explored by Aloimonos and Shulman [17]. The idea was then extended to SVMs by Nilsback and Caputo [18] (DAS). The resulting method showed remarkable performances on visual object recognition applications. In this paper we will use a variant of the original DAS algorithm, using Least Square-SVM (LS-SVM) instead of SVM. We made this choice because LS-SVM provides as output a better confidence estimate (this point will be addressed more thoroughly in section III-B).

## III. Audio-Visual Cue Integration

Due to the fundamental difference in how audio and visual information is acquired and processed, it is reasonable to assume that they provide different kinds of information. Thus, we expect that by combining them through an integration scheme, we will achieve a better performance, namely higher classification performance and higher robustness.

As it was reviewed in the previous section, several authors suggested different methods to combine information derived from different cues. They can all be re-conducted to one of these three approaches: *high-level*, *mid-level* and *low-level* integration [19]. We tested a LS-SVM-based high-level integration scheme on the task at hand, namely the Discriminative Accumulation Scheme (DAS, [18]). In this method each single cue first generates a set of hypotheses on the correct label of the test image, and then those hypotheses are combined together so to obtain a final output. The algorithm is revised in section III-B. Another possible strategy is mid-level integration, where the features are merged during the classification step. To this end, we designed a new class of kernels, the Multi-Cue Kernel (MCK), that accepts as input different cues while building a unique optimal separating hyperplane. This new kernel is described in details in section III-C. Finally we decided to use a low level integration scheme. This kind of approach is based on concatenating existing feature vectors in a new one, so in a sense it builds a new representation. It is questionable if this approach can solve the robustness problem because if one of the cues gives misleading information it is possible that the new feature vector will be adversely affected. A description of this strategy is given in section III-D.

### A. Least Square Support Vector Machines

Assume $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$, with $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \{-1, 1\}$, is a set of samples drawn from an unknown probability distribution. We want to find a function $f(\mathbf{x})$ such that $sgn(f(\mathbf{x}))$ best determines the category of any future sample $\mathbf{x}$ drawn from the same distribution. In Least-Squares Support Vector Machine (LS-SVM) we construct a linear model $f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b$, where $\phi(\mathbf{x})$ is a non-linear function that maps the data in a fixed feature space. However, rather than specifying the feature space directly, it can be implied by a kernel function $K(\mathbf{x}, \mathbf{x}')$, giving the inner product between the images of vectors in the feature space, i.e. $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$. A common kernel function is the isotropic Gaussian kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||x - x'||^2) \tag{1}$$

that will be used in our experiments.

The solution is found minimizing a regularized least-squares loss function [5]. This approach is similar to the well-known formulation of Support Vector Machines. The difference is that the loss function is the least square and it does not induce a sparse solution. On the other hand it is possible to write the leave-one-out error in closed form [20]. This is known to be approximately an unbiased estimator of the classifier generalization error [21]. This is useful to find the best parameters for the learning (e.g. $\gamma$ in (1)). Another advantage is the fact that the outputs converges to the conditional in-class probabilities [22]. This is opposed to SVM outputs that instead do not carry any information of the confidence on a predicted label [22]. For this reason LS-SVM are more suited, with respect to SVM, for approaches that combine the outputs of different classifiers, using their confidence estimate.

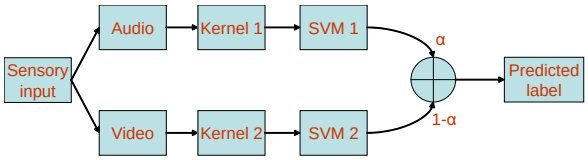


Fig. 1. A schematic illustration of the high-level cue integration approach.

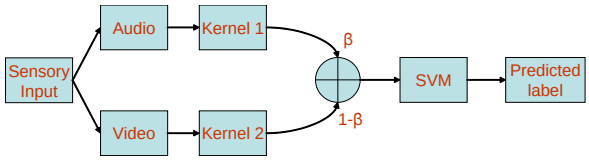


Fig. 2. A schematic illustration of the mid-level cue integration approach.

### B. High-Level Cue Integration

High-level cue integration methods start from the output of two or more classifiers, dealing with complementary information. Each of them produces an individual hypothesis about the object to be classified. All those hypotheses are then combined together, so to achieve a consensus decision. In this paper we applied this integration strategy using the Discriminative Accumulation Scheme (DAS, [18]). It is based on a weak coupling method called accumulation [11], which does not neglect any cue contribution. The DAS main idea is that information from different cues can be summed together. Figure 1 illustrates schematically the basic idea behind this approach.

Suppose we are given $M$ object classes and for each class, a set of $N_j$ training vector data $\{I_i^j\}_{i=1}^{N_j}$, $j = 1, \dots M$. For each vector, we extract a set of $P$ different cues:

$$T_p = T_p(I_i^j), \quad p = 1 \dots P \tag{2}$$

so that for an object $j$ we have $P$ new training sets $\{T_p(I_i^j)\}_{i=1}^{N_j}, j = 1, \dots M, p = 1 \dots P$. In the case of multi-modal vector data, each new training set will correspond to a different modality. In case of unimodal vector data, each new training set will correspond to different unimodal cues. Of course it is also possible to consider the case where, from multi-modal training data, one extracts different unimodal cues from each sensor channel. This case will not be considered in this paper. For each new training set, we train a LS-SVM. Kernel functions may differ from cue to cue and model parameters can be estimated during the training step via cross validation. Given a test vector $\hat{I}$ and assuming $M \geq 2$, for each single-cue LS-SVM we compute the distance from the separating hyperplane:

$$D_j(p) = \sum_{i=1}^{m_j^p} \alpha_{ij}^p y_{ij} K_p \left( T_p(I_i^j), T_p(\hat{I}) \right) + b_j^p. \tag{3}$$

After collecting all the distances $\{D_j(p)\}_{p=1}^{P}$ for all the $j$ objects $j = 1, \dots, M$ and the $p$ cues $p = 1, \dots, P$, we classify the vector $\hat{I}$ using the linear combination:

$$j^* = \underset{j=1}{\overset{M}{\operatorname{argmax}}} \{ \sum_{p=1}^{P} a_p D_{j(p)} \}, \quad a_p \in \Re^+. \tag{4}$$

The coefficients $\{a_p\}_{p=1}^{P}$ are evaluated via cross validation during the training step.

### C. Mid-Level Cue Integration

Combining two cues at a median level means that the different features descriptors are kept separated, but they are integrated in a single classifier generating the final hypothesis. Figure 2 illustrates schematically the approach. To implement this approach we developed a scheme based on multi-class LS-SVM with a Multi Cue Kernel $K_{MC}$. This new kernel combines different features extracted from the vector data. The Multi Cue Kernel is a Mercer kernel, as positively weighted linear combination of Mercer kernels are Mercer kernels themselves [23]:

$$K_{MC}(\{T_p(I_i)\}_p, \{T_p(I)\}_p) = \sum_{p=1}^{P} a_p K_p(T_p(I_i), T_p(I)). \tag{5}$$

In this way it is possible to perform only one classification step, identifying the best weighting factors $a_p$ while determining the optimal separating hyperplane; this means that the coefficients $a_p$ are guaranteed to be optimal.

### D. Low-Level Cue Integration

To combine two or more feature vectors it is possible to start from the descriptors, and to combine them together in a new representation. In this way the cue integration does not directly involve the classification step. This fusion strategy is called low-level. Figure 3 shows schematically the basic idea behind this approach. For the problem at hand we chose feature concatenation as the fusion approach: two feature vectors $f_i$ and $c_i$ are concatenated into a single feature vector $v_i = (f_i, c_i)$ that is normalized to one and is then used for classification. In this fusion strategy the information related to each cue is mixed without a weighting factor that allows to control the influence of each information channel on the final recognition result. In general terms a drawback of this method is that the dimension of the feature vector increases as the number of of cues grows, implying longer learning and recognition times, greater memory requirements and possibly curse of dimensionality effects. Moreover, it is not always possible to use the low-level integration approach: there are features that have a variable number of vector's elements per input data, while some other have a defined number of them. Due to their intrinsic nature, the first one asks for specialized classification algorithms and it is not possible to combine them together with vectors of the second kind. Furthermore, information from different modalities is not always acquired at the same time (synchronicity issue), which opens the question on how to create a unique representation from these inputs.

## IV. Experiments

This section describes our experimental evaluation of the three different cue integration methods. We first describe

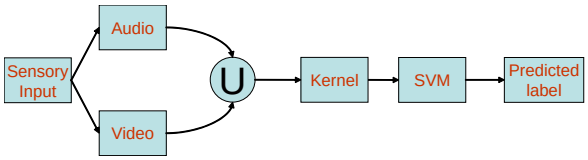


Fig. 3. A schematic illustration of the low-level cue integration approach.

the databases used (section IV-A). Then we describe the method employed for audio data processing (section IV-B) and the algorithm used for video data processing (section IV-C). Section IV-D reports the results obtained and discusses our findings.

### A. Databases

We conducted experiments on audio-visual data relative to two different object categories: dogs and guitars. The audio-visual data were artificially generated from existing audio-only and vision-only databases, as described below.

The visual data for the categories dogs and guitars were taken from the Caltech Dataset [1]. For each object category, we preliminary selected several corresponding background classes, containing natural scenes or various distracting objects. For both object categories, we conducted a set of preliminary experiments in order to select the most challenging background, between those available. The goal of this procedure was to create a difficult task for the vision-only recognition algorithm. This should make it easier to evaluate the impact on performance of the cue integration strategies, and it should allow for comparison between them. This set of experiments resulted in the background 'site' selected for the object category dogs, and for the background 'road' selected for the object category guitars. Figure 4 shows exemplar images of these visual classes with the relative backgrounds.

The audio dataset consisted of a large number of audio clips, manually collected from the Internet, corresponding to the visual categories dogs and guitars. The audio background noise class contains recordings of road traffic noise and pedestrian zone noise. Each audio file (object/background) is randomly associated with an image (object/background) without any repetition. In this way we attempted to reproduce the natural coupling between audio and video signals.

### B. Audio Data Processing

Realistic audio data in general is characterized by its strong amplitude modulation content, i.e., signal energy exhibits a large variance when observed with a time-constant of about 30 ms. To capture the modulation structure of the sounds, signals were first decomposed into 17 different spectral "ERB" bands from about 50 Hz to about 3800 Hz with a spectral width of one ERB unit that resembles the logarithmically scaled sensitivity of human and animal auditory systems. Log-scaled signal amplitudes within each band were analyzed with a second spectral decomposition of 1 s long windows that characterized the time-scale of the amplitude modulations from 2 Hz to 30 Hz within this spectral band. Hence, the original time-domain audio signal was transformed into the 3-dimensional representation of the "amplitude modulation spectrogram" [24] with dimensions time, frequency and modulation frequency. For each 1 s long temporal window modulation intensity values at $13 \times 29 = 377$ points in the frequency/modulation-frequency plane are derived, which comprise the set of features from which feature selection can pick the best ones.

Classification was performed using the cue integration methods described in the previous section. In order to evaluate properly the impact of multi-modality on the final performance, we also performed audio-only recognition experiments, using a LS-SVM algorithm with a Gaussian kernel, trained to discriminate between audio samples containing only background noise (road traffic or pedestrian zone sounds) and samples containing an acoustic category (such as a dog, or a guitar). Input features for the LS-SVM were taken from the 377-dimensional amplitude modulation spectrogram representation.

### C. Video Data Processing

To learn object models, we use the method described in [25]. The method starts by extracting interest regions using the Kadir & Brady [26] feature detector. After their initial detection, selected regions are cropped from the image and scaled down to $11 \times 11$ pixel patches, represented using the first 15 DCT coefficients (not including the DC). To complete the representation, 3 additional dimensions are concatenated to each feature, corresponding to the $x$ and $y$ image coordinates of the patch, and its scale respectively. Therefore each image $I$ is represented using an unordered set $F(I)$ of 18 dimensional vectors. The algorithm learns a generative relational part-based object model, modeling appearance, location and scale. Each part in a specific image $I_i$ corresponds to a patch feature from $F(I_i)$. It is assumed that the appearance of different parts is independent, but this is not the case with the parts' scale and location. However, once the object instances are aligned with respect to location and scale, the assumption of part location and scale independence becomes reasonable. Thus a 3-dimensional hidden variable $C = (C_l, C_s)$, which fixes the location of the object and its scale, is used. The model's parameters are discriminatively optimized using an extended boosting process. For the full derivation of the model and further details, we refer the reader to [25]. The LS-SVM is trained with a Gaussian Kernel.

### D. Results

Table I shows the results of 10 random training/testing splits ($75\%/25\%$) on audio-only, vision-only and audio-visual data, for the two object categories dogs and guitars and the three cue integration schemes. For all experiments we used the Gaussian kernel. The algorithms' parameters were selected though leave one out cross validation, using the closed formula for LS-SVM [5]. The weights for DAS

[1] Available at `http://www.vision.caltech.edu/archive.html`


Fig. 4. Sample images from the datasets. Object images appear on the left, background images on the right.

TABLE I

RECOGNITION RESULTS FOR THE CATEGORIES GUITARS AND DOGS. RESULTS ARE REPORTED FOR UNIMODAL DATA AND FOR THE THREE DIFFERENT CUE INTEGRATION SCHEMES

| | Guitars | Dogs |
|---|---|---|
| Audio | $91.11 \pm 1.64$ | $88.76 \pm 2.52$ |
| Video | $96.20 \pm 1.06$ | $87.64 \pm 2.14$ |
| Low level | $98.06 \pm 0.53$ | $89.91 \pm 2.02$ |
| MCK | $98.10 \pm 0.70$ | $91.20 \pm 2.49$ |
| DAS | $99.03 \pm 0.40$ | $93.42 \pm 1.71$ |

were found using the leave one out predicted outputs from each model and selecting the weights with smaller error.

As a first comment, we see that using multi-modal information always yields better performances than using only one sensory channel, for both object categories. For the category guitars, the gain in performance goes from a minimum of +1.86%, achieved by the low-level approach with respect to the vision-only classifier, to a maximum of +7.92%, achieved by the DAS approach with respect to the audio-only classifier. Similar performances are achieved also on the category dogs: the gain in performance goes from a minimum of +1.15%, obtained by the low-level approach with respect to the audio-only classifier, to a maximum of +5.78%, obtained by DAS with respect to the vision-only method. A second remark is about the three cue integration method. We see that, for both classes, DAS outperforms the other two approaches. With respect to the low-level approach, the gain in performance goes from a +0.97% for the category guitars, to a + 3.51% for the category dogs. With respect to the mid-level approach, the increase in performance goes from a + 0.93% for the category guitars, to a + 2.22% for the category dogs. It is interesting to note that the greater improvement is always observed for the category dogs. This category has the lowest unimodal performances between the two categories. This result thus suggests that the way multi-modal cues are combined together becomes more important for challenging cases.

## V. CONCLUSIONS AND FUTURE WORK

This paper presented a comparative evaluation of three different cue integration schemes for audio-visual object category detection. We considered respectively a high-level, a mid-level and a low-level integration scheme. Features were extracted from the multi-modal data using state of the art approaches. All the integration schemes were based on the least square support vector machine algorithm. Experiments were performed on data, artificially generated, representing two object categories, dogs and guitars, presenting different audio and visual characteristics. Results showed that the high-level cue integration approach performs better than the other two proposed methods. This is in agreement with a consistent body of literature in the neuroscience.

This work can be developed in many way. First, the approaches should be tested on real audio-visual data. We are currently collecting an audio-visual database for gender classification. Using this data will permit to evaluate the performance of our methods on noisy inputs, and it will also allow to explore the synchronicity issue, which has been purposefully neglected in this paper. Second, we would like to evaluate the methods when more than two cues are integrated. This could be done by extracting several unimodal cues from each sensor inputs. Finally, we want to test several confidence measures for the DAS algorithm, and compare the effectiveness of SVM-based method for confidence estimate as opposed to probabilistic classifiers. Future work will be devoted to these issues.

## REFERENCES

[1] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and System*, 2004.

[2] G. A. Calvert, C. Spence, and B. E. Stein, *The handbook of the multisensory processes*. MIT Press, Cambridge MA., 2004.

[3] D. Burr and D. Alais, "Combining visual and auditory information," *Progress in brain research*, vol. 155, 2006.

[4] L. Cahill, F. Ohl, and H. Scheich, "Alteration of auditory cortex activity with a visual stimulus through conditioning. a 2-deoxyglucose analysis," *Neurobiol Learn and Mem*, vol. 65, pp. 213–222, 1996.

[5] N. Cristianini and J. S. Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

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

[7] M. M. Murray, S. Molholm, C. M. Michel, D. J. Heslenfeld, W. Ritter, D. C. Javitt, C. E. Schroeder, and J. J. Foxe, "Grabbing your ear: rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment," *Cereb Cortex*, vol. 15, pp. 963–974, 2005.

[8] P. Lakatos, C.-M. Chen, M. N. O'Connell, A. Mills, and C. E. Schroeder, "Neuronal oscillations and multisensory interaction in primary auditory cortex," *Neuron*, vol. 53, pp. 279–292, 2007.

[9] C. Kayser and N. K. Logothetis, "Do early sensory cortices integrate cross-modal information?" *Brain Struct Funct*, vol. 212, pp. 121–132, 2007.

[10] A. Fillbrandt, M. Deliano, and F. W. Ohl, "Audiovisual category transfer - an electrophysiological study in rodents." in *Proceedings of the 7th Meeting of the German Neuroscience Society/31st Gttingen Neurobiology Conferenc*, 2007, pp. T28–9B.

[11] J. Clark and A. Yuille, *Data fusion for sensory information processing systems*. Kluwer Academic Publisher, 1990.

[12] S. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, "Generating est. of class. conf. for a case-based spam filter," in *Proc. ICCBR'05*.
[13] J. Platt, "Probabilistic outputs for SVMs and comparisons to regularized likelihood methods," in *Adv. in Large Margin Classifiers*, 2000.
[14] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class class. by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, 2004.
[15] J.-J. Kim, B.-W. Hwang, and S.-W. Lee, "Retrieval of the top n matches with support vector machines," in *Proc. ICPR'00*.
[16] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, 1985.
[17] J. Aloimonos and D. Shulman, *Integration of Visual Modules: an Extension of the Marr Paradigm*. Academic Press, 1989.
[18] M. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," in *Proceedings of the International conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 578–585.
[19] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.
[20] R. Rifkin and R. Lippert, "Notes on regularized least-squares," Massachusetts Institute of Technology, Cambridge, MA, Tech. Rep. CBCL Paper #268/AI Technical Report #2007-019, May 2007.
[21] A. Luntz and V. Brailovsky, "On estimation of characters obtained in statistical procedure of recognition (in russian)," *Techicheskaya Kibernetica*, vol. 3, 1969.
[22] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *The Annals of Statistics*, vol. 32, pp. 56–134, March 2004.
[23] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. CUP, 2000.
[24] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1593–1602, 1994.
[25] A. Bar-Hillel and D. Weinshall, "Efficient learning of relational object class models," *IJCV*, 2007.
[26] T. Kadir and M. Brady, "Saliency, scale and image description," *IJCV*, 2001.