Object Category Detection using Audio-visual Cues

Luo Jie[1,2], Barbara Caputo[1,2], Alon Zweig[3],
Jörg-Hendrik Bach[4], and Jörn Anemüller[4]

[1] IDIAP Research Institute, Centre du Parc, 1920 Martigny, Switzerland
[2] Swiss Federal Institute of Technology in Lausanne(EPFL), 1015 Lausanne, Switzerland
[3] Hebrew university of Jerusalem, 91904 Jerusalem, Israel
[4] Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany
{jluo,bcaputo}@idiap.ch zweiga@cs.huji.ac.il {joerg-hendrik.bach, joern.anemueller}@uni-oldenburg.de

Abstract. Categorization is one of the fundamental building blocks of cognitive systems. Object categorization has traditionally been addressed in the vision domain, even though cognitive agents are intrinsically multimodal. Indeed, biological systems combine several modalities in order to achieve robust categorization. In this paper we propose a multimodal approach to object category detection, using audio and visual information. The auditory channel is modeled on biologically motivated spectral features via a discriminative classifier. The visual channel is modeled by a state of the art part based model. Multimodality is achieved using two fusion schemes, one high level and the other low level. Experiments on six different object categories, under increasingly difficult conditions, show strengths and weaknesses of the two approaches, and clearly underline the open challenges for multimodal category detection.

Key words: Object Categorization, Multimodal Recognition, Audio-visual Fusion

1 Introduction

The capability to categorize is a fundamental component of cognitive systems. It can be considered as the building block of the capability to think itself [1]. Its importance for artificial systems is widely recognized, as witnessed by a vast literature (see [2, 3] and references therein). Traditionally, categorization has been studied from an unimodal perspective (with some notable exceptions, see [4] and references therein). For instance, during the last five years the computer vision community has attacked the object categorization problem by *(a)* developing algorithms for detection of specific categories like cars, cows, pedestrian and many others [2, 3]; *(b)* collecting several benchmark databases and promoting benchmark evaluations for assessing progresses in the field. The emerging paradigm from these activities is the so-called 'part-based approach', where visual categories are modeled on the basis of local information. This information is then used to build a learning based algorithm for classification. Both probabilistic and discriminative approaches have been used so far with promising results.

Still, an algorithm aiming to work on an autonomous system cannot ignore the intrinsic multimodal nature of categories, and the multi sensory capabilities of the system.

For instance, we do recognize people on the basis of their visual appearance and their voice. Linen can be easily recognized because of its distinctive textural visual and tactile properties; and so forth. Biological systems combine information from all the five senses, so to achieve robust perception (see [5] and references therein).

In this paper we propose an audio-visual object category detection algorithm. We consider categories like vehicles (cars, airplanes), instruments (pianos, guitars) and animals (dogs, cows). We assume that the category has been localized, and we focus on how to integrate together effectively the two modalities. We represent visual information using a state of the art part based model (section 2, [3]). Audio information is represented by a discriminative classifier, trained on biologically motivated spectral features (section 3.1, [6]). Following results from psychophysics, we propose to combine the two modalities with a high level fusion scheme that extends previous work on integration of multiple visual cues (section 3.2, [7]). Our approach is compared with single modality classifiers, and with a low level integration approach. Experiments on six different object categories, with increasing level of difficulty, show the value of our approach and clearly underline the existing challenges in this domain (section 3.3).

2 Vision Based Category Detection

In this section we present the chosen vision based category detection algorithm (section 2.1) and experiments showing its strengths and weaknesses (section 2.2).

2.1 Visual Category Detection

To learn object models, we use the method described in [3]. The method starts by extracting interest regions using the Kadir & Brady (KB) [8] feature detector. After their initial detection, selected regions are cropped from the image and scaled down to 11×11 pixel patches, represented using the first 15 DCT (Discrete Cosine Transform) 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 [3].

2.2 Experiments

We used an extensive dataset of six categories (airplanes, cars, cows, dogs, guitars and pianos). They present different type of challenges: airplanes and cars contain relatively

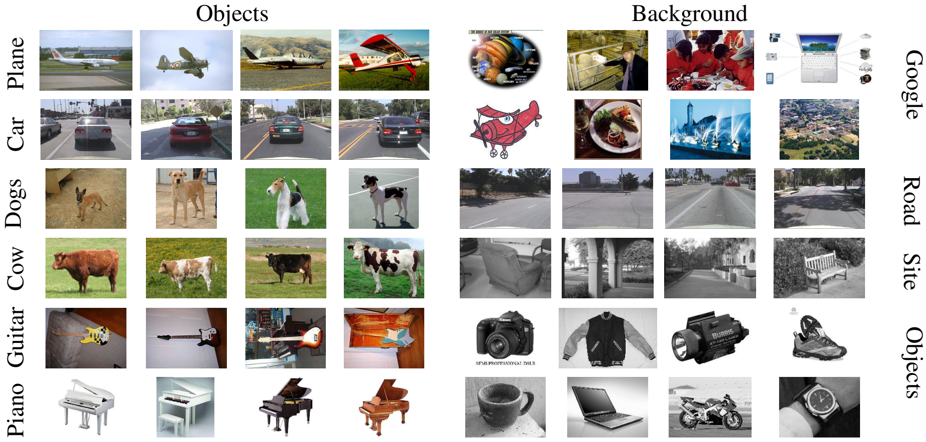


Fig. 1. Sample images from the datasets. Object images appear on the left, background images on the right.

small variations in scale and location, while cows and dogs have a more flexible appearance and variations in scale and locations. The category images and the background classes were collected from standard benchmark datasets (Caltech Datasets[5] and PASCAL Visual Challenge[6]). For each category we have several corresponding testing backgrounds, containing natural scenes or various distracting objects. Images from the six categories and the background groups are shown in Figure 1. Each category was trained and tested against different backgrounds. Each experiment was repeated several times, with randomly generated training and test sets. Table 1 presents the average results, for different categories and varying backgrounds. These numbers can be compared with those reported in [3], and show that the method delivers state of the art performance. We then run some experiments to challenge the algorithm. Namely, we reduced the training set to roughly 1/3 for some categories (airplanes, cars; results reported in Figure 2) and, for all categories, we collected new test images containing strong occlusions, unusual poses and high categorical variability. Exemplar challenging images are shown in Figure 2. We used the learnt models to classify these challenging images. These results are also reported in Figure 2. We see that, under these conditions, performance drops significantly for all categories. Indeed, these results seem to indicate that the part-based approach might suffer when different categories share similar visual part (dogs and cows sharing legs, cars and airplanes sharing wheels), or when the variability within a single category is very high, as it is for instance for grand pianos and upright pianos, or classic and electric guitars. It is worth stressing that these considerations are likely to apply to *any* part-based visual recognition method. Thus, our multimodal approach for overcoming these issues is of interest for a wide variety of algorithms.

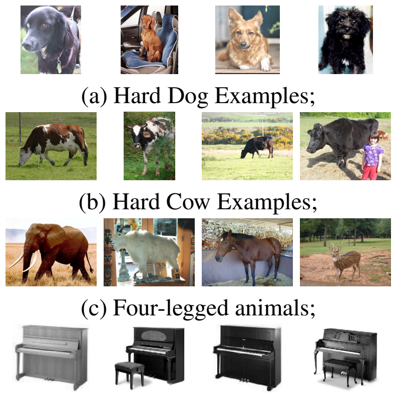
[5] Available at `http://www.vision.caltech.edu/archive.html`
[6] Available at `http://www.pascal-network.org/challenges/VOC/`

Object	Background	FNR	FPR	ERR	Background	FNR	FPR	ERR
Airplanes	Google	2.05	0.81	1.46	Road	1.90	4.96	3.62
Cars	Google	5.70	0.41	2.25	Road	11.39	0.73	3.69
Cows	Site	19.70	2.22	6.91	Road	9.09	0.18	1.14
Dogs	Site	17.00	13.89	15.53	Road	1.90	0.55	0.91
Guitars (Electrical)	Google	9.36	7.59	8.49	Objects	3.69	1.68	2.75
Pianos (Grand)	Google	18.89	1.41	4.23	Objects	6.67	6.56	6.60

Table 1. Performance of our visual category detection algorithm on six different objects on various background. $False\ Negative\ Rate(FNR) = \frac{num.\ of\ false\ neg.}{num.\ of\ pos.\ instances}$, $False\ Positive\ Rate(FPR) = \frac{num.\ of\ false\ pos.}{num.\ of\ neg.\ instances}$ and $Error\ Rate(ERR) = \frac{num.\ of\ false\ prediction}{total\ num.\ of\ instances}$ are reported separately.

Object	Background	Visual		
		FNR	FPR	ERR
Airplane*	Road	26.29	13.89	19.97
Car*	Road	38.02	1.91	13.54
Cow$^\circ$	Site	78.33	-	78.33
Dog$^\circ$	Site	27.00	-	27.00
Piano†	Google	58.48	-	58.48
Guitar†	Google	16.00	-	16.00

*: reduced number of training samples;
$^\circ$: learnt models of cows & dogs to detect new test images with occlusions and strange pose;
†: learnt models of grand pianos & electrical guitars to detect upright piano and classical guitar respectively.



(a) Hard Dog Examples; (b) Hard Cow Examples; (c) Four-legged animals; (d) Upright Piano

Fig. 2. Performance of the visual category detection algorithm on various difficulty examples. Some exemplary images are shown on the right of the table.

3 Audio-visual Category Detection

This section presents our multi-modal approach to object category detection. We begin by illustrating the sound classification method used (section 3.1). We then illustrate our integration method (section 3.2) and show with an extensive experimental evaluation the effectiveness of our approach (section 3.3).

3.1 Audio Category Detection

Real-world audio data is characterized in particular by two properties, spectral characteristics and modulation characteristics. Spectral characteristics are obtained by decomposing the signal into different spectral bands, typically using "Bark-scaled" frequency bands that approximate the spectral resolution of the human ear. Here, we use 17 Bark bands ranging from about 50 Hz to 3800 Hz. Within each spectral band, information about the signal is encoded in changes of spectral energy across time, so-called amplitude modulations (2 Hz to 30 Hz). Grouping both properties in a single diagram, we obtain the "amplitude modulation spectrogram" (AMS, [9]), a 3-dimensional signal representation with dimensions time, (spectral) frequency and modulation frequency. Each 1s long temporal window is represented by $17 \times 29 = 493$ points in frequency/modulation-frequency space. Audio category detection [6] is performed by linear SVM classification based on a subset of the 493 AMS input features, trained to discriminate between audio samples containing only background noise (e.g., street) and samples containing an audio category object (e.g., dog) embedded in background noise at different signal-to-noise ratios (from +20 dB to -20 dB).

3.2 Audio-visual Category Detection

This section provides a short description of our cue integration scheme. Many cue integration methods have been presented in the literature so far. For instance, one can divide them in *low level* and *high level* integration, where the emphasis is on the level at which integration happens [4]. In *low level* integration, information is combined before any use of classifiers or experts. In *high level* approaches, integration is accomplished by an ensemble of experts or classifiers; on each prediction, a classifier provides a hard decision, an expert provides an opinion. In this paper, we will investigate methods from both approaches.

High Level Integration There are several methods for fusing multiple classifiers at the decision level [10], such as voting, sum-, product-rule, etc. However, voting could not be easily applied on our setup, since it requires an odd number of classifiers for a two class problem, and more for a multi-class problem. Here we use an extension of the *discriminative accumulation scheme (DAS)* [7]. The basic idea is to consider the margin outputs of any discriminative classifiers (e.g. AdaBoost and SVMs) as a measure of the confidence of the decision for each class, and accumulate all the outputs obtained for various cues with a linear function. The binary class version of the algorithm could be described into two steps:

1. *Margin-based classifiers:* These are a class of learning algorithms which take as input binary labeled training examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$ with $\boldsymbol{x}_i \in \chi$ and $\boldsymbol{y}_i \in \{-1, +1\}$. Data are used to generate a real-valued function or hypothesis $f : \chi \to \Re$, with f belonging to some hypothesis space F. The margin of an example $\boldsymbol{x}$ with respect to f is $f(\boldsymbol{x})$, which is determined by minimizing: $\frac{1}{m}\sum_{i=1}^{m} L(y_i f(\boldsymbol{x}))$, for some loss function: $L : \Re \to [0, \infty]$ Different choices of the loss function L and different algorithms for minimizing the equation over some hypothesis space lead to various well studied learning algorithms such as Adaboost and SVMs.
2. *Discriminative Accumulation:* After all the margins are collected $\{f_j^p\}_{p=1}^{P}$, for all the P cues, the data $\boldsymbol{x}$ is classified using their linear combination:

$$J = \operatorname{sgn}\left(\sum_{p=1}^{P} w_p f_j^p(\boldsymbol{x}_p)\right).$$

The original DAS method considered only SVM as experts, used multiple visual cues for training and determined the weighting coefficients via cross validation. Here we generalize the approach in many respects: we take two different large margin classifiers (SVM and AdaBoost) as experts, we train each expert on a different modality, and we determine the weights $\{w_p\}_{p=1}^{P}$ by training a single-layer artificial neural network (ANN) on a validation set.

A drawback of the original DAS algorithm is that the accumulation function is linear, thus the method is not able to adapt to the special characteristics of the model. For example, one sensor might be suddenly affected by noise, or detect a novel input. Here we will assume that if one sensor is very confident about the presence of a category (i.e. margin above a certain threshold), it is highly probable that this sensor is correct. We thus introduce a threshold before the accumulation, so that if the margin output value of one classifier is larger than the threshold, we will take it directly as the decision.

Low Level Integration The *low level* fusion is also known as feature level fusion. Features extracted from data provided by different sensors are combined. In case of audio and visual feature vectors, the simple concatenation technique could be employed, where a new feature vector can be built by concatenating two feature vectors together. There are a few drawbacks to this approach: the dimensionality of the resulting feature vector is increased, and the two separate feature vectors must be available at the same time (synchronous acquisition). Due to the second problem, the *high level* integration is usually preferred in the literature for audio-visual fusion [4].

The visual feature vectors are built by concatenating all the P feature descriptors. Each feature consists of a 20-dimensional vector including [3]: the 18 dimensional vector representing each image (see section 2.1), plus a normalized mean of the feature and a normalized logarithm of the feature variance. The training set is then normalized to have unit variance in all dimensions, and the standard deviations are stored in order to allow for identical scaling of the test data. Finally, the visual feature vector is concatenated with audio feature vectors, and a linear SVM is trained for the detection.

3.3 Experiments

Experimental Setup We evaluated our multi-modal approaches with three series of experiments. Our audio dataset contains a large number of audio clips, manually collected from the internet, corresponding to the six visual categories as well as some other objects. The audio background noise class contains recordings of road traffic and pedestrian zone noise. All audio models were trained with the same background noise but different object sounds, on several combinations of training and test sets, randomly generated. Then each audio file (object/background) was randomly associated with an image (object/background) without repetitions. Audio and visual data were collected separately, and their association was somehow arbitrary. Thus, we repeated the experiments at least 1,000 times, for each setup, so to prevent "lucky" cases. We compared our result on fusion with those obtained by single cues, reporting average results. For DAS, we experimented with the linear and non-linear (i.e. with an additional threshold for high-confidences input) approaches. However, we did not find significant differences between them. Thus we only report results obtained using the linear method.

	Object	Background	Audio			Visual			Fusion		
			FNR	FPR	ERR	FNR	FPR	ERR	FNR	FPR	ERR
High-level	Airplane	Road	7.53	11.41	9.71	1.55	2.99	2.36	1.36	1.52	1.45
	Cars	Road	6.73	3.22	4.20	11.33	0.71	3.67	1.51	0.70	0.93
	Cows	Site	4.62	0.47	1.49	19.97	2.21	6.97	2.40	0.62	1.10
	Dog	Site	2.09	0.36	1.27	16.78	13.79	15.37	0.36	0.62	0.48
	Piano	Google	7.98	0.09	1.36	18.87	1.41	4.21	1.75	0.31	0.54
	Guitar	Google	7.53	0.09	3.16	9.32	7.57	8.29	0.48	0.30	0.38
Low-level	Airplane	Road	8.12	6.19	7.03	4.63	9.57	7.40	3.49	2.31	2.83
	Cars	Road	7.52	1.35	3.06	8.27	1.46	3.35	3.67	0.19	1.16
	Cows	Site	5.37	0.03	1.47	16.74	6.44	9.20	5.13	0.00	1.39
	Dog	Site	1.96	0.01	0.99	16.93	22.91	19.76	1.83	0.01	0.97
	Piano	Google	8.50	0.00	1.37	21.56	0.81	4.16	7.50	0.00	1.21
	Guitar	Google	4.15	0.06	2.15	18.5	1.45	10.17	3.80	0.02	1.96

Table 2. Results of each separate audio and visual cues and detection performance of both the high- and low-level integration scheme on six different objects.

Experiments with Clean Data Table 2 reports the FNR., FPR., and ERR. for different objects, using *high-* and *low-level* fusion schemes. For each object, we performed experiments using various backgrounds. For space reasons we report here only a representative subset. Results show clearly that, for all objects and both fusion schemes, recognition improves significantly when using multiple cues, as opposed to single modalities. Regarding the comparison between the two fusion approaches, it is important to stress that, due to the different classification algorithms, and the differences in statistics in the training data for the different classes, it is not straightforward how to compare the performance of the two fusion schemes. Still, the high-level scheme seems to obtain overall lower error rates, compared to the low-level approach.

Experiments with Difficult/Noisy Data We tested the robustness of our system with respect to noisy cues or difficult sensory inputs. First, we showed the effects of including audio cues for improving the system performance when there are not enough training examples (see section 2.2); results are reported in Table 3. Then, we used the models trained on clean data and test them on various difficult images (see section 2.2). These results are also shown in Table 3. Finally, the systems were tested against noisy audio inputs. The performance of the audio classifier was deliberately decreased by adding varying amount of street noise on the test object audio (SNR $\in [-20db, 20db]$), and including a varying amount of audio files generated by other objects [7] as part of the test background noise (from 25% to 1% of the total number of testing background samples). Average results on two selected examples, car (less training examples, road background) and dog (site background), are shown in Figure 3.

With respect to the high-level and low-level fusion methods, we see that the low level approach seems to be more robust to noise (Table 3 and Figure 3). This might be

[7] Sounds of other animals, e.g. bear, horse, in case of experiments on cows and dogs, and sounds of artificial objects, e.g. phone, helicopters, in case of vehicles and instruments.

	Object	Background	Audio			Visual			Fusion		
			FNR	FPR	ERR	FNR	FPR	ERR	FNR	FPR	ERR
High.	Airplane (Less)	Road	8.72	11.90	10.33	26.24	13.89	19.97	6.17	7.62	6.90
	Cars (Less)	Road	6.80	3.31	4.46	38.02	1.91	13.54	3.69	1.76	2.38
Low.	Airplane (Less)	Road	5.36	9.65	7.76	3.78	29.16	18.01	2.76	6.96	5.12
	Cars (Less)	Road	7.50	1.57	3.21	14.08	10.01	11.15	4.83	0.50	1.70

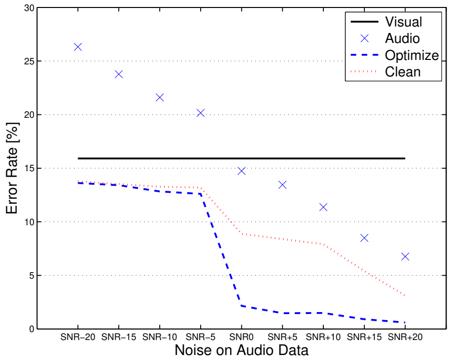
Table 3. Performance of the multimodal system suffering from less visual training examples.

Object	Background	*High-level*			*Low-level*		
		Audio	Visual	Fusion	Audio	Visual	Fusion
Cows (Hard)	Site	7.24	78.33	7.27	7.54	77.02	7.92
Dog (Hard)	Site	2.69	27.00	2.32	2.56	29.24	2.54
Piano (Upright)	Google	7.21	58.48	16.35	8.43	39.39	8.41
Guitar(Classical)	Google	7.14	16.00	6.40	6.04	29.80	5.93

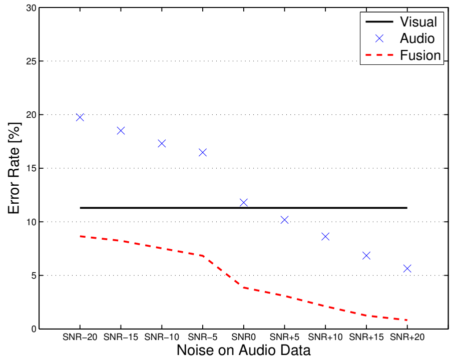
Table 4. Performance of the multimodal system in presence of difficult test examples with occlusions, unusual poses and high categorical variability. Since background images were not used during test, only the false negative error rates (FNR) are reported.

due to the nature of the two algorithms, as for the low level fusion method the error rate is linked to the lower error rate between the two cues. The high-level fusion scheme instead weights the confidence estimates from the two sensory channels with coefficients learned during training. Thus, if one modality was weighted strongly during training, but is very noisy during test, the high-level scheme will suffer from that.

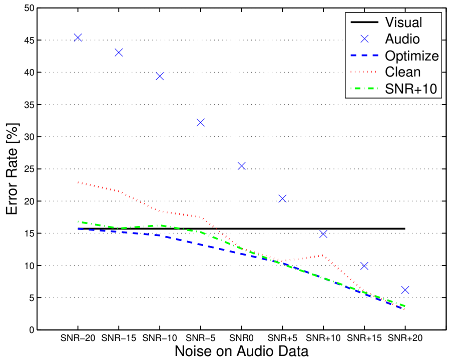
Experiments with Missing Audio An important issue when working on multimodal information processing is the synchronicity of the two modalities, i.e. both audio and visual input must be perceived together by the system. However, unlike the multimodal person authentication scenario [4], in real-world cases the two inputs may not always be synchronized, e.g. a dog might be quiet. We tested our system in the case where some of the object samples were not accompanied by audio. To simplify the problem, we only considered the case where roughly 50% of the object samples are "silent". We considered three different ways to tackle the problem (see Figure 4, caption), and we optimized our system using a validation set under the same setup. Figure 4 reports the average results on the categories car (road background) and dog (site background). We can see that the performance still improves significantly when the missing audio inputs were represented using zero values (roughly the same as results reported in Table 2). For the other two scenarios, the performance on cars still grows, while the performance on dogs drops because the system was biased toward the audio classifier when the visual classifier did not have high accuracy. However, the system was always better than using the visual algorithm alone.



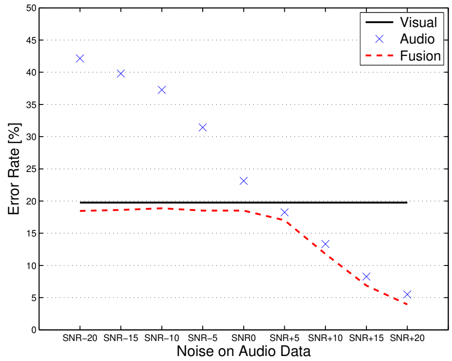
(a) Cars, High-level Fusion.



(b) Cars, Low-level Fusion.



(c) Dogs, High-level Fusion.



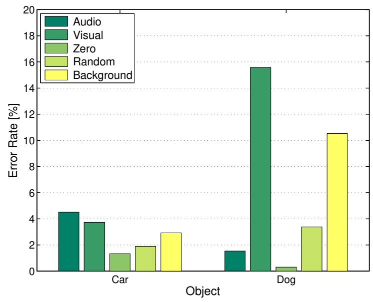
(d) Dogs, Low-level Fusion.

Fig. 3. Performance of the multimodal system in the presence of different level of corrupted audio inputs. For high-level fusion, the accumulating weights were found using different criteria: the weights were determined using clean audio and visual data through previous experiments (Clean), determined using data at current test noisy level (Optimize), or determined using data at fixed noisy level (e.g. SNR+10).

4 Discussion and Conclusions

This paper presented a multimodal approach to object category detection. We considered audio and visual cues, and we proposed two alternative fusion schemes, one high-level and the other low-level. We showed with extensive experiments that using multiple modalities for categorization leads to higher performance and robustness, compared to uni-modal approaches.

This work can be developed in many ways. Our experiments show that the high-level approach might suffer in case of noisy data. This could be addressed by using adaptive weights, related to the confidence of the prediction for each modality. Also, we estimate confidences using the distance from the separating hyperplane, but other solutions should be explored. We also plan to extend our model to a hierarchical representation as in [11]. Finally, these experiments should be repeated on original audio-visual



Three ways for representing the missing audio input:
Zero: the input confidences of the audio classifier equal zero, if the audio is missing; thus only the visual classifier will be considered.
Random: the input confidences of the audio classifier equal randomly generated numbers with a zero mean and standard deviation equals one, if the audio is missing;
Background: based on the assumption that the environmental noise was always presented, the test images were associated with a random background audio if the audio input is missing.

Fig. 4. Performance of the multimodal system under asynchronous test conditions.

data, so to better address the issues of synchronicity and sound-visual localization. Future work will focus on these issues.

Acknowledgments This work was sponsored by the EU integrated project DIRAC (Detection and Identification of Rare Audio-visual Cues, http://www.diracproject.org/) IST-027787. The support is gratefully acknowledged.

References

1. Pfeifer, R., Bongard, J.: How the body shapes the way we think. MIT Press (2006)
2. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. Int. J. Comput. Vision **71**(3) (2007) 273–303
3. Bar-Hillel, A., Weinshall, D.: Efficient learning of relational object class models. Int. J. Comput. Vision (2007) in press.
4. Sanderson, C., Paliwal, K.K.: Identity verification using speech and face information. Digital Signal Processing **14**(5) (2004) 449–480
5. Burr, D., Alais, D.: Combining visual and auditory information. Progress in Brain Research **155** (2006) 243–258
6. Schmidt, D., Anemüller, J.: Acoustic feature selection for speech detection based on amplitude modulation spectrograms. In: 33rd German Annual Conference on Acoustics. (2007)
7. Nilsback, M.E., Caputo, B.: Cue integration through discriminative accumulation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2004) 578–585
8. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vision **45**(2) (2001) 83–105
9. Kollmeier, B., Koch, R.: Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. J. Acoust. Soc. Am. **95**(3) (1994) 1593–1602
10. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits and Systems Mag. **6**(3) (2006) 21–45
11. Zweig, A., Weinshall, D.: Exploiting object hierarchy: Combining models from different category levels. In: IEEE 11th International Conference on Computer Vision. (2007) 1–8