The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Audio/Video Fusion for Objects recognition

Loic Lacheze Yan Guo Ryad Benosman Bruno Gas Charlie Couverture

Abstract— **In mobile robotics applications, pattern and object recognition are mainly achieved relying only on vision. Several other perceptual modalities are also available such as, touch, hearing or vestibular proprioception. They are rarely used and can provide valuable additional information within the recognition tasks. This article presents an analysis of several methods of fusion of perceptual and auditory modalitites. It relies on the use of a perspective camera and a microphone on a moving object recognition problem. Experimental data are also provided on a database of audio/visual objects including cases of visual occlusions and audio corruptions.**

I. INTRODUCTION

The integration of multiple sensory modalities has been defined as a necessity to ease learning, simplify the calculation of various features, and to allow a robust perception of the environment [9]. The complementarity of modalities simplifies several reputated difficult tasks, and can be applied to a wide variety of applications and domains : vision [21], objects recognition [8], robot localization [15], etc. Existing work merging audio and video are mainly related to applications within man-machine interactions [16], [19]. Existing systems usually combine audio and video with other sensors such as lasers to detect faces [12], and use audio to localize sound sources within scenes [6] sometimes also using microphone arrays [4], [5].

From a theoretical point of view a large part of the existing merging methods involve mutual information between modalities [13], [11] or are based on statistical techniques [14], [7]. The aim of this paper is to focus on the case of audio-vision fusion, and to inquire on what should really be extracted from images and sounds to reach an easy and efficient merging of both modalities. A new method of extracting features from images using an entropic sampling is used, it allows a better description of observed scenes and a non constrained feature extraction. Practical experiments will focus on the recognition of moving objects each emitting sounds during motion. These experiments provide a practical framework to study the interaction of image and sound in tasks of objets recognition. The main important points being the ability to understand the amount of necessary correlations between both modalities and how they are related to each other. Fusion results are presented in cases of severe alterations of the objects' database by adding visual occlusions and audio corruptions.

UPMC Univ Paris 06, UMR 7222, F-75005, Paris, France, 4 Place Jussieu, BP 173, 75252 Paris Cedex 05, FRANCE. ryad.benosman@upmc.fr

II. SYSTEM ARCHITECTURE

A. Visual system

In the last few years, the problem of extracting features from images has received growing attention. The majority of existing methods use derivatives approaches and rely on local image patches as basic features [1]. Recently, bag-of-features representations have become popular, they are geometry free, based purely on characterizing the statistics of local patch appearances [10]. The idea behind the method is to extract a set of local image patches which are sampled and assigned a metric description. The resulting quantified descriptors give an implicit distribution description space that can be quantified using different methods. Most of the existing work differ mainly according to the way patches are sampled and then described. They are generally selected using keypoints SIFT based approaches [3]. Codebooks are then produced using k-means and agglomerative clustering [10]. Most of the cited techniques consider partial information from scenes mainly distributed around maximal gradient points, which limits the robustness of visual loops. The used method is driven by the idea that all the information contained in images is useful. It is based on a dense multilayer decomposition of the image driven by the quantiy and homogeneity of the information contained within subpatches.

B. Entropic decomposition of images

An efficient decomposition must produce an optimal and possibly a unique partitioning of images. In addition it would be interesting to produce less patches, but of variable size so that they can cover homogeneous texture zones. In order to generate patches, a quadtree-like algorithm is set up. Quadtree algorithms cut recursively images into subimages and so on. Starting for the initial image, each subimage is cut into four equal subimages. The idea is to use the same principle, but at the contrary of the regular quadtree approach, the division of subimages will be driven by an entropy measure. The idea is to cut a subimage at the location were the difference of the quantity of information between possible subimages is minimal. All scales will then contain valuable information as the sampling is ensuring that information is equally distributed in subimages. This quantity of information is given for an image I by: $H(I) = -\sum_{c=0}^{c=255} Occ(I=c)\log P(I=c)$ with $Occ(I=c)$ the number of times the pixel value c appears in I, $P(c)$ is the probability of appearance of the grey value c within I.

To estimate the optimal point minimizing the variance of the distance between the information contained in the four subimages of I, the principle of integral images introduced in [2] is used.

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

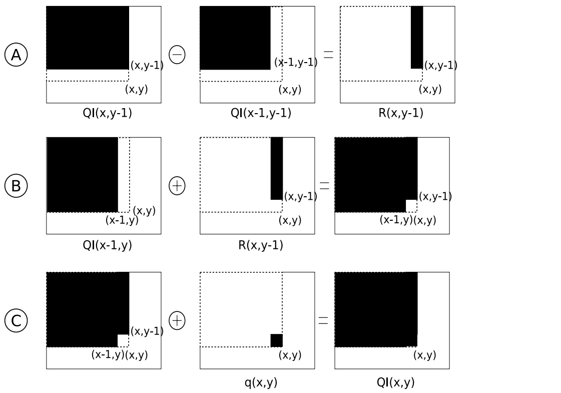

Fig. 1. Computation of $QI(x,y)$ using $QI(x-1,y-1)$, $QI(x,y-1)$, $QI(x-1,y)$ and $q(x,y)$. In (A) The computation of $R(x,y-1)$ and in (B) and (C) the final computation of $QI(x,y)$.

Let $q(i,j)$ be the quantity of information of a pixel $I(i,j)$ with $q(i,j) = \log(P(I(i,j)))$. We set the integral information of $I(x,y)$ as:

$$QI(x,y) = \sum_{i \leq x, j \leq y} q(i,j) \tag{1}$$

This sum is computed in one iteration on the whole image or subimage considered. We set $R(x,y-1)$ the integral quantity of information on the row x of height $y-1$. The principle of computation is presented in Fig. 1(A).

$$R(x,y-1) = QI(x,y-1) - Q(x-1,y-1). \tag{2}$$

Finally the integral quantity of information for a (x,y) (see Fig. 1(B)(C)) is given by:

$$QI(x,y) = QI(x-1,y) + R(x,y-1) + q(i,j). \tag{3}$$

Once QI is computed the variance value within each pixel becomes implicit. It is important to compute the mean value of the quantity of information contained within the four subimages. In the case where I is of size $m \times n$ we have for a cutting position $(x = m/2, y = n/2)$: $QI_m = QI(m,n)/4$. As shown in Fig. 2, the quantity of information of each zone is:

$$\begin{cases} QI_{11}(x,y) = QI(x,y), \\ QI_{12}(x,y) = QI(m,y) - QI(x,y), \\ QI_{21}(x,y) = QI(x,n) - QI(x,y), \\ QI_{22}(x,y) = QI(m,n) - QI_{21} - QI_{12} + QI_{11}. \end{cases} \tag{4}$$

Finally the optimal (x,y) position is the one minimizing the following sum of differences:

$$\exists(x,y)/min_{x,y}(\sum_{a=1,b=1}^{a=2,b=2}(QI_m - QI_{ab}(x,y))^2). \tag{5}$$

To illustrate the stability of the decomposition, Fig. 2 (*below*) presents the selected patches for a simple object translated between the two images. One may check the stability of the method by noticing that in both cases patches cover the same zones despite the important translation of the considered object. This is an expected result, the amount of information being the same between in both images, the decomposition generates the same patches.

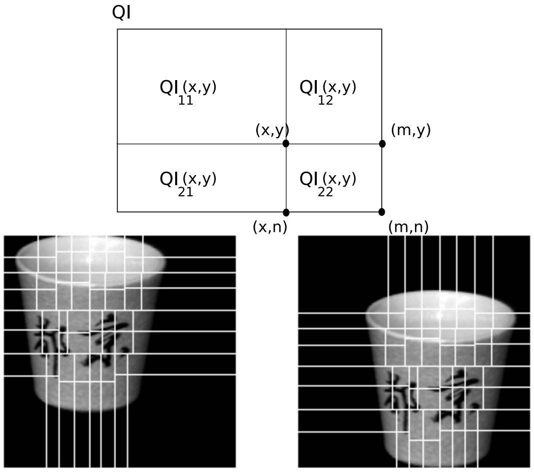

Fig. 2. *Up:* Computation of the quantity of information contained in the four subimages of QI for a cutting position located at (x,y). *Down:* Optimal generation of patches for a translated object. The object is covered with the same patches, covering the same areas and providing an equal decomposition of the image.

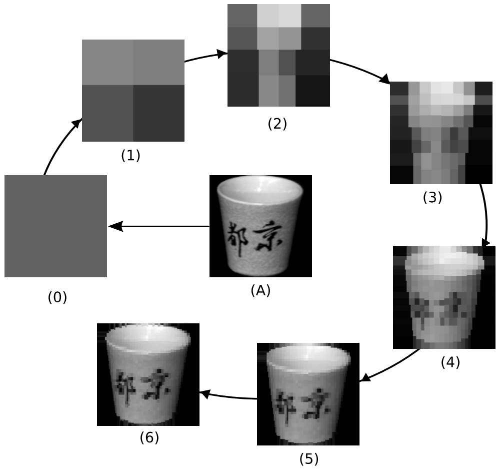

Fig. 3. *Example of codification of a scene using the optimal method for every layer of Fig. (A). Each patch contains its mean color in grey level.*

C. Description of Patches and local decision

Let $F_z(I)$ be the function allowing the decomposition of the image using the entropic sampling I into several textured patches :

$$F_z(I) = z_0, z_1, ..., z_n \text{ with } I = \bigcup_{i=0}^{n} z_i \tag{6}$$

Extracted patches are described according to their texture using a method as shown in [25]. Let $T = h_{z_0}, h_{z_1}, ..., h_{z_n}$ be the set containing all texture descriptors of patches z_i of I. The idea is to sample T to reduce the number of descriptors to $m \leq n$, m being adjusted for all images to 20. We then add to T a metric function expressed by $dist(h_{z_i}, h_{z_j})$ and a reference texture patch h_{ref}. The reference patch is set to a patch containing a single color, corresponding to a uniform area. In a second stage all the representation of patches contained in T are compared to h_{ref} and sorted, from the less to the more textured. The set T_s corresponding to the ordered set T becomes :

$$T_s = h_{ref}, h'_{z_0}, h'_{z_1}, ...h'_{z_n} \tag{7}$$

with :

$$dist(h_{ref}, h'_{z_i}) \leq dist(h_{ref}, h'_{z_j}) \text{ if } i < j \tag{8}$$

The mahalanobis distance is used as a metric function. At this point, T_s still contains too much information, it is then sampled into m equal areas. For each area, only the median patch is selected. Extracted patches are then identified within a predefined codebook V. Codebooks are generated by selecting within a database of images a wide variety of subimages to ensure a wide coverage of possible patches, in this paper it is set to 32 elementary patches. A codified image I_{cod_i} is then a vector containing all codebook indexes of its extracted patches. Comparing two codified images I_{cod_i} and I_{cod_j} is then given by :

$$Comp(I_{cod_i}, I_{cod_j}) = \sum_{i=1}^{m} (I_{cod_i} - I_{cod_j})^2 \tag{9}$$

Substracting images introduce a spatial dimension to the comparison, as due to the optimal patch decomposition two close images will theoretically have the same decomposition.

III. AUDIO SYSTEM

A. Predictive modelization of audio frames

Sounds are non-stationary and highly redundant signals which require to be divided into successive short frames and then encoded.

This encoding allows to generate a set of coefficients representative of the short-term spectrum. The coding that we use is based on the model of the cochlea of Patterson [20]. In the model of Patterson, the bandwidth of each cochlear filter is described by an equivalent rectangular bandwidth (ERB). Each filter models the signal present at the output of a nerve in the cochlea. For N ERB filters we obtain N signals. These signals are segmented successively into frames of length L on which we calculate the energy terms:

$$\mathbf{x}_k = [x_1^k, x_2^k, \dots x_N^k] \text{ with } x_{i,i=1\dots N}^k = \sum_{q=1}^{L} |y_i^k(q)|^2$$

where x_i^k represents the component of the code vector $\mathbf{x}_k$ calculated on the frame k and the $y_i(q)^k$ represents the $q = 1 \dots, L$ samples of the output signal of the i^{th} filter.

It would be of great interest to use frame duration of 40ms as it could allow an easy synchronization of audio and video streams since the video sampling period is 40ms. Unfortunatly, the auditory scenes are subject to spectral stationarity constraints which lead us to use frames of 10ms to 20ms duration. Given an audio sampling frequency of 48kHz we used overlapped frames of 480 samples to obtain exactly 8 audio frames corresponding to one 40ms video frame.

B. Local decision

The approach we propose for processing audio data is based on a technique already used in applications of speaker recognition: the neural predictive modeling of sound sources [17]. This technique makes it possible to estimate a distance between an unknown audio source and a set of reference sources. For that it takes into account the spectral components of the signal (in our case the coefficients calculated on N filters) at the same time as their dynamic evolution when the signal is not stationary.

Let $\mathbf{x}_k = [x_1^k, x_2^k, \dots x_N^k]$ be the cochlear coding vector corresponding to the k^{th} audio frame where N is the number of coefficients. For a M classes problem (M objects to recognize), M neural networks are trained so as to associate every two consecutive coding vectors $\mathbf{x}_{k-2}$, $\mathbf{x}_{k-1}$ to the corresponding next vector $\mathbf{x}_k$. Once the learning of the M networks is completed, they all are representative of the M sound objects. During the recognition phase, an unknown signal coming from the audio sensor is first segmented and then coded by the cochlear model. Every sequences of two coding vectors $\mathbf{x}_{k-2}$, $\mathbf{x}_{k-1}$ are then presented as input of the M networks (see Fig. 4) so that M prediction errors are computed between the real next frame $\mathbf{x}_k$ and the predicted next frame $\hat{\mathbf{x}}_k$: $\varepsilon_k = \hat{\mathbf{x}}_k - \mathbf{x}_k$. The local decision consists in labeling the unknown frame with the class of the network giving the minimal prediction error. The neural networks used are multilayer perceptrons with one hidden layer.

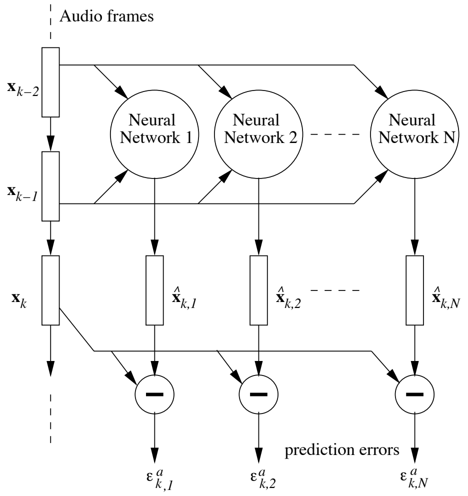

Fig. 4. Predictive architecture with neural networks.

IV. FUSION AND DECISION STAGE

The recognition of sound objects follows two stages: a stage of local decision made at the frame level followed by a stage of overall decision. Two kind of local decisions are made. First by fusing errors computed on the audio and on the video frames, making then the local decision, and also by merging decisions already made independently at the audio level and the video level. The overall decision consists in affecting a class to the whole sequence from all the local assumptions.

A. Local decision

In what follows we note $\varepsilon_{k,i}^a$ as the prediction error computed from the outputs of the network representing the sound class $i, i \in \{1, \dots, M\}$ of the audio frame k. This term is therefore homogeneous to the inverse of the i^{th} class-membership likelihood of the frame k. We note $\varepsilon_{k,i}^v$ the equivalent error term provided by the vision algorithm. We compared 4 methods of fusion between audio and video data for the local decision algorithm.

1) Decision without fusion (LD1): Both audible and visual modalities are treated separately.The overall decision is made from all local decisions $\{c_k^a, c_k^v\}$ estimated separately on each audio and video frame by minimizing the error [24]:

$$\left\{ \begin{array}{l} c_k^a = \arg\min_{i=1,M}\{\varepsilon_{k,i}^a\}, \\ c_k^v = \arg\min_{i=1,M}\{\varepsilon_{k,i}^v\}. \end{array} \right. \tag{10}$$

2) Extended Decision without fusion (LD2): This fusion algorithm works as the previously defined *LD*1 algorithm, adding the second minimum prediction errors, i.e, errors defined by :

$$\left\{ \begin{array}{l} c_k^{a(2)} = \arg\min_{i=1,\dots,M, i\neq c_k^a}\{\varepsilon_{k,i}^a\}, \\ c_k^{v(2)} = \arg\min_{i=1,\dots,M, i\neq c_k^v}\{\varepsilon_{k,i}^v\}. \end{array} \right. \tag{11}$$

that is from all local decision $\{c_k^a, c_k^{a(2)}, c_k^v, c_k^{v(2)}\}$.

3) Decision with simple fusion (LD3): For each frame k, a local decision is made by choosing the class minimizing the error, audio or video. Each frame is therefore labelized according to the following rule:

$$c_k = \arg\min\{\varepsilon_k^a, \varepsilon_k^v\} \text{ with } \varepsilon_k^{\{a,v\}} = min_{i=1,\dots,M}\{\varepsilon_{k,i}^{\{a,v\}}\} \tag{12}$$

Note that this method needs to previously normalize the ε_k^a and ε_k^v errors before making the final local decision :

$$\varepsilon_k^m \longleftarrow \frac{1}{\sigma^m}\left(\varepsilon_k^m - \overline{\varepsilon_k^m}\right) \tag{13}$$

$\overline{\varepsilon_k^m}$ being the mean error value and σ^m the standard deviation, all of them computed over all the video and audio frames composing the audio-visual sequence.

4) Decision with weighted fusion (LD4): Among the simpliest methods of merging, one is the fusion by weighted decision between audio and video errors:

$$c_k = \arg\min\{\alpha_k^a \varepsilon_k^a, \alpha_k^v \varepsilon_k^v\} \text{ with } \varepsilon_k^m = min_{i=1,\dots,M}\{\varepsilon_{k,i}^m\} \tag{14}$$

The weighting criterion is based on the ratio between the minimum error noted $\varepsilon_k^{m(1)}$ and the second minimum error $\varepsilon_k^{m(2)}$ of both audio and visual modalities, as defined previously:

$$\alpha_k^{m\in\{a,v\}} = \frac{\varepsilon_k^{m(1)}}{\varepsilon_k^{m(2)}}. \tag{15}$$

As underlined before, the $\varepsilon_k^{m(1)}$ should be first normalized. The α^m are quality factors defined such as to emphasize the modality which seems more suitable : the lowest the ratio $\varepsilon_k^{m(1)}/\varepsilon_k^{m(2)}$ is, the most likely the m modality decision is.

B. Global decision

In the previous section we presented 4 fusion methods of local decision. We have thus a set of local decisions $c_{k,k=1,\dots,K}$ where K represents the total number of frames. We propose two ways to build a global decision. The first algorithm (*GD*1) does not consider local decisions but instead computes a global error by summation of all the local minimum errors, making then a decision by global error minimization. This algorithm will give us a baseline recognition rate for comparing results obtained from fusion algorithms. The second algorithm (*GD*2) makes a global decision from all the local decisions.

- (*GD*1) decision (*global error*) or Global Decision 1:

$$c = \arg\min_{i,i=1,\dots,M}\left\{\sum_{k=1}^{K}\varepsilon_{k,i}\right\} \tag{16}$$

- (*GD*2) decision (*simple majority*) or Global Decision 2:

$$c = \arg\max_i\{|U_i|\}, \tag{17}$$

where U_i is the set of all the frames of class $i, i \in \{1,\dots,M\}$ and $|U_i|$ its cardinal.

Because of their simplicity of implementation, these two types of decision algorithms are commonly used in problems of sequence classification, at least when not using statistical methods (such as hidden Markov models, for example).

V. EXPERIMENTS

The experiments presented have been conducted to test the complementarity of audio and video information in a task of moving objects recognition with visual occlusions and noise disturbances.

A. Experimental conditions

We have built a database of audiovisual recordings of 28 *toys* moving in front of a uniform background. Each object emits a particular noise due to its movement on the ground and eventually issuing a characteristic sound. Fig. 5 presents a global view of these objects. One can distinguish sub-classes within shape, colour and sounds emited. Objects 1-5 are spherical and emit very similar rolling sounds. Objects 13-19 have wheels, similar shapes and dimensions but different colors. Objects 26-28 have dissimilar shapes and emit caracteristic sounds that are proper to them (artificial sounds non related to displacement).

For the purposes of our experimentation, we made three recordings for each of the objects, leading to the formation of three databases as following:

1) *B*1: Learning database allowing to adapt audio and video models of the objects;
2) *B*2: Test database without any occlusion. We test the performances of models in normal conditions.
3) *B*3: Test database with visual occlusions during motion (see Fig. 5).
4) *B*4: Test database with auditory occlusions. This database is generated using B2 on which we added a white gaussian noise artificially.

All objects go through the visual scene horizontally from left to right for *B*1 and *B*3, and from right to left for the test base *B*2 and *B*4. A small panel placed between the camera and the object is used to introduce visual occlusions, the panel covers approximately 30% of the visual scene (see Fig. 5 (*below*)).

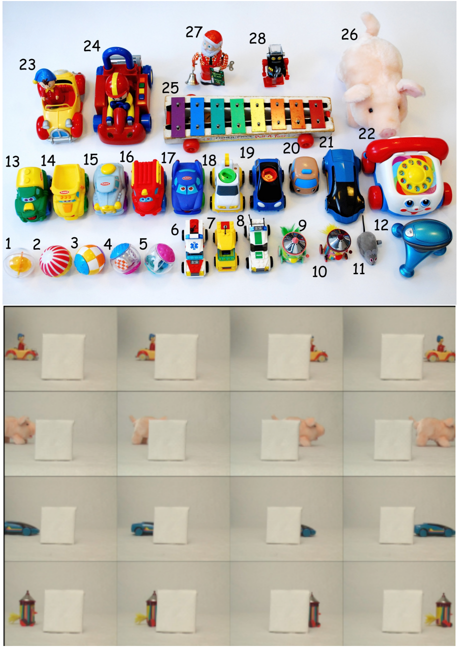

Fig. 5. *Up*: Global view of the objects composing the database. *Down*: Images extracted from 4 sequences with visual occlusion.

The white gaussian noise added to $B2$ with a $-10dB$ signal to noise ratio in order to obtain $B4$ is present 50% of time.

Digital audio signals are provided by the camera's microphones at a rate of 48kHz. In order to precisely caracterize spectral informations, we use 13 coefficients by frame, which seems a good compromise between low-dimension vectors and minimal information-loss. The multi-Layer Perceptrons we use have 26 inputs and 13 outputs. Simulations have shown that the best performance in learning were obtained with predictive networks of 6 hidden cells. 5000 learning iterations were enough to obtain the algorithm convergence of the 28 models.

B. Experimental results

1) Classification without fusion: After extracting and learning both the audio-visual features from the 28 objects of $B1$, we conducted a first test of recognition on $B2$, $B3$ and $B4$ without fusion: the objects were therefore recognized separately by the audio and visual loops using criterion $GD1$ (Eq. 16). Table I shows the results, on $B2$, $B3$ and $B4$.

In the case of objects of $B2$ (free from any perturbations the recognition score are respectively of 96,43% and 71,43%. They tend to show that it is more difficult to try recognize the objects using separate modalities. In fact many small objects of the database emit similar sounds introducing legitimate confusions. It seems that the visual extraction of features is robust enough as it succesfully identified 27 objects with a high recognition rate of 96,43%.

Visual occlusions introduced in $B3$ lower drastically the recognition rate from 96,43% to 57,14%. The stability of the algorithm is seriously affected on the while sequence, the recognition is only possible on the non occulted frames that do not succed in allowing a stable rate along acquisition. The results are the same in the case of audio recognition rates. The important pertubation of the audio signal drop the results from 71,43% to 10,71%.

TABLE I

RECOGNITION TESTS ON $B2$, $B3$, $B4$ WITHOUT FUSION USING $GD1$

Databases	Video	Audio
B2	96,43%	71.43%
B3	57.14%	71.43%
B4	96,43%	10.71%

2) Classification with fusion: We will now compare the fusion strategies in order to show the complementarity of audio-visual features by testing fusion algorithms $LD1$, $LD4$ on databases $B2$, $B3$ and $B4$.

TABLE II

TESTS RECOGNITION ON $B2$, $B3$, $B4$ WITH $GD1$ GLOBAL DECISION AND WITH LOCAL FUSION $LD1$ TO $LD4$ AND $GD2$ GLOBAL DECISION

Decision →	$GD1$	$GD2$			
Bases		$LD1$	$LD2$	$LD3$	$LD4$
B2	89.29%	100%	96.43%	82.14%	100%
B3	75.00%	78.57%	75.00%	71.43%	78.57%
B4	14.29%	96.43%	60.71%	67.86%	96.43%

The first column of Table II can be directly compared to the results of Table I. In both cases the global decision algorithm is $GD1$. On $B2$ the fusion of modalities score is adjusted as the median of both the 96.43% for vision and 71,43% for audio. At the contrary in the presence of visual occlusion ($B3$), the recognition rate increases after fusion to 75% from initial scores of 57.14% and 71.43%. This illustrates the complementarity of the modalities when using adpated fusion. In $B4$ the deterioration of the audio modalitiy has a very low impact on the overall score due to the high rates provided by vision.

The following columns present the results obtained using the second global fusion algorith $GD2$ and the four local decisions $LD1 - LD4$. Without occlusions (B2), the fusion allows to increase the recognition rates up to 100% in the case of LD1 and LD4. In the case of visual occlusions, GD2 increases recognition rates at a level very similar to GD1. Algorithms LD1 and LD4 always show the highest results. Finally in the case of the audio occlusions, GD2 appears to be highly more robust than GD1 as the rates increase from 14.29% to 96.43%. These results show that a global decision computed from local decisions GD2 is a much more powerfull fusion strategy than a global decision relying on a summation of local errors GD1.

VI. CONCLUSIONS AND FUTURE WORKS

We presented in this article a series of non-statistical fusion algorithms showing the interest of the multimodal approach in the tasks of perception in robotics, particularly in connection with the recognition of objects in motion. We proposed to adress two modalities: the visual and audio in the framework of recognition objects emitting sound while moving. We have proposed a vision algorithm allowing the recognition of objects from image sequences using a bag of words technique. The audio data processing was carried out by more conventional methods as used in speech recognition. The experiments presented showed that it is possible to get better scores for recognition by merging multimodal information rather than adressing each modality separately. In order to illustrate this, we realized a base of 28 dynamic objects accessible online on our website (www.isir.fr). Our longer term goal is to show that multisensory perception is not the addition of modalities treated separately, while each can in their own way solve specific problems in recognition. The graal being the discovery of new common feature spaces that will allow a more direct solving of fusion perceptual modalities problems.

REFERENCES

[1] Frederic Jurie and Bill Triggs, ”Creating Efficient Codebooks for Visual Recognition”, *International Conference on Computer Vision*, 2005.

[2] Viola, P. and Jones, M. (2001). ”Rapid object detection using a boosted cascade of simple features”, pp 1332-1338,San Diego.(2007)

[3] Lowe, D., ”Distinctive image features from scale-invariant keypoints”, *International Journal of Computer Vision (IJCV)*,91–110, 2004.

[4] S. Argentieri, P. Danes, P. Souères and P. Lacroix. ”An Experimental Testbed for Sound Source Localization with Mobile Robots using Optimized Wideband Beamformers” In pp 909–914, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2005

[5] S. Argentieri, P. Danes, P. Souères and P. Lacroix. ”Modal Analysis Based Beamforming for Nearfield or Farfield Speaker Localization” in Robotics, pp 866–871, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006

[6] M. J. Beal, H. Attias and N. Jojic. ”Audio-Video Sensor Fusion with Probabilistic Graphical Models”, In pp 736–752, *ECCV 2002*.

[7] M. Beal, N. Jojic and H. Attias. ”A graphical model for audiovisual object tracking”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp 828–836, 2003.

[8] Beltran-Gonzalez C. and G. Sandini. ”Visual attention priming based on crossmodal expectations”. In: *Intelligent Robots and Systems*, pp 1060–1065, 2005. (IROS 2005).

[9] R. A. Brooks, C. Breazeal, R. Irie, C. Kemp, M. Marjanovic, B. Scassellati, and M. Williamson. ”Alternative essences of intelligence”. In *Proceedings of the 15th AAAI*, AAAI Press, pp 961–968, 1998.

[10] Filliat, D. ”A visual bag of words method for interactive qualitative localization and mapping”. In: *International Conference on Robotics and Automation (ICRA)*, 2007.

[11] J. W. Fisher and Trevor Darrell. ”Informative Subspaces For Audio-Visual Processing: High-Level Function From Low-Level Fusion” (2002)

[12] , A. Haasch, S. Hohenner, S. Huwel, M. Kleinehagenbrock, S. Lang, I Topsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer. ”BIRON, The Bielefeld Robot Companion”. In: *Proc. Int. Worshop on Advances In Service Robots*, Stuttgart, 2004.

[13] J. Hershey and J. R. Movellan, ”Audio vision: Using audiovisual synchrony to locate sounds”, in *Advances in Neural Information Processing Systems* 12. S. A. Solla, T. K. Leen and K. R. Muller (eds.), pp 813–819. MIT Press., 2000.

[14] Thomas Hofmann. ”Probabilistic Latent Semantic Analysis”. In: *Proc. of Uncertainty in Artificial Intelligence*. 1999.

[15] Irie, R. ”Mutlimodal Sensory Integration for localization” in *Humanoid Robot*. In Proceedings of Second IJCAI Workshop on Computational Auditory Scene Analysis (CASA’97), IJCAI-97.

[16] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer. ”Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot”. In *Proc. Int. Conf. on Multimodal Interfaces*, pp 28–35, Vancouver, Canada, November 2003. ACM.

[17] Melouk, A. and Gallinary, P. ”A discriminative neural predictive system for speech recognition”. *ICASSP*, Vol. 2, pp 533–536 (1993).

[18] Mermelstein, P. ”Distance measures for speech recognition, psychological and instrumental”. *Pattern Recognition and Artificial Intelligence*, pp 374–388, 1976.

[19] K. Nakadai, H. G. Okuno, and H. Kitano, ”Robot recognizes three simultaneous speech by active audition”, in *Proc. of IEEE International Conference on Robotics and Automation (ICRA 2003)*.

[20] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C.Zhang, and M. H. Allerhand, ”Complex sounds and auditory images,” In *Auditory Physiology and Perception*, (Eds.) Y Cazals, L. Demany, K.Horner, Pergamon, Oxford, 1992, pp 429–446.

[21] J. Peskin and B. Scassellati. ”Image Stabilization through Vestibular and Retinal Feedback”. In R. Brooks, ed. *Research Abstract*, MIT Artificial Intelligence Laboratory. 1997.

[22] Rumelhart, D.E. and Hinton, G.E. and Williams, R.J. ”Learning representations by back-propagating errors”. *Nature*, Vol. 323, pp 533–536 (1986)

[23] I. Ulrich and I. Nourbakhsh. ”Appearance Based Place Recognition for topological Localization”. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp 1023–1029, 2000.

[24] Ming Liu, Ziyou Xiong, Stephen M. Chu, Zhenqiu Zhang and Thomas S.Huang. ”Audio visual word spotting”. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[25] Osada, R., Funkhouser, T., Chazelle, B., and Dobkin, D. (2002). ”Shape distributions”. *ACM Transactions on Graphics*