Isolated Spanish Digit Recognition based on Audio-Visual Features

Gonzalo D. Sad, Lucas D. Terissi and Juan C. Gómez

Lab. for System Dynamics and Signal Processing, Universidad Nacional de Rosario, Argentina
CIFASIS-CONICET, Rosario, Argentina
{sad, terissi, gomez}@cifasis-conicet.gov.ar

Abstract. The performance of classical speech recognition techniques based on audio features is degraded in noisy environments. The inclusion of visual features related to mouth movements into the recognition process improves the performance of the system. This paper proposes an isolated word speech recognition system based on audio-visual features. The proposed system combines three classifiers based on audio, visual and audio-visual information, respectively. An audio-visual database composed by the utterances of the digits (in Spanish language) is employed to test the proposed system. The experimental results show a significant improvement on the recognition rates through a wide range of signal-to-noise ratios.

Keywords: Speech recognition, audio-visual speech features, Hidden Markov Models

1 Introduction

In recent years, significant research efforts have been devoted to the development of Multimodal Human Computer Interfaces (HCIs) that try to imitate the way humans communicate with each other, which is inherently a multimodal process, in the sense that, for the transmission of an idea, not only is important the acoustic signal but also the facial expressions and body gestures [4]. For instance, a significant role in spoken language communication is played by lip reading. This is essential for the hearing-impaired people, and is also important for normal listeners in noisy environments to improve the intelligibility of the speech signal. Audio Visual Speech Recognition (AVSR) is a fundamental task in HCIs, where the acoustic and visual information (mouth movements, facial gestures, etc.) during speech are taken into account. Several strategies have been proposed in the literature for AVSR [7][6], where improvements of the recognition rates are achieved by fusing audio and visual features related to speech. As it is expected, these improvements are more notorious when the audio channel is corrupted by noise, which is a usual situation in speech recognition applications. These strategies usually differ in the way the audio and visual information is extracted and

combined, and the AV-Model employed to represent the audio-visual information. These approaches are usually classified according to the method employed to combine (or fuse) the audio and visual information, *viz.*, feature level fusion, classifier level fusion and decision level fusion [2].

In feature level fusion (early integration), audio and visual features are combined to form a unique audio-visual feature vector, which is then employed for the classification task. This strategy is effective when the combined modalities are correlated, since it can exploit the covariance between the audio and video features. This method requires the audio and visual features to be exactly at the same rate and in synchrony, and usually performs a dimensionality reduction stage, in order to avoid large dimensionality of the resulting feature vectors. In the case of classifier level fusion (intermediate integration), the information is combined within the classifier using separated audio and visual streams, in order to generate a composite classifier to process the individual data streams [5]. This strategy has the advantage of being able to handle possible asynchrony between audio and visual features. In decision level fusion (late integration), independent classifiers are used for each modality and the final decision is computed by the combination of the likelihood scores associated to each classifier [3]. Typically, these scores are fused using a weighting scheme defined based on the reliability of each unimodal stream. This strategy does not require strictly synchronized streams.

In this paper an isolated digit recognition system based on audio-visual features is proposed. This system is based on the combination of early and late fusion schemes. In particular, acoustic information is represented by mel-frequency cepstral coefficients, and visual information is represented by coefficients related to mouth shape. The efficiency of the system is evaluated considering noisy conditions in the acoustic channel. The proposed system combines three classifiers based on audio, visual and audio-visual information, respectively, in order to improve the recognition rates through a wide range of signal-to-noise ratios (SNRs). A Spanish audio-visual database is employed to test the proposed system. The experimental results show that a significant improvement is achieved when the visual information is considered.

The rest of this paper is organized as follows, the audio, visual and audio-visual features used for each classifier are described in section 2 together with the database used for the experiments. The proposed classifiers and the early integration strategy are analyzed in section 3. A general description of the proposed system using the late fusion scheme is given in section 4. In section 5 experimental results are presented, where the performance of the proposed strategy is analyzed. Finally, some concluding remarks and perspectives for future work are included in section 6.

2 Audiovisual Database and Features

In order to evaluate the proposed speech recognition system an audio-visual database was compiled. This database consists of videos of 16 speakers facing

the camera, pronouncing a set of ten words 20 times, in random order. These words correspond to the Spanish utterances of the digits from zero to nine. The videos were recorded at a rate of 60 frames per second with a resolution of 640×480 pixels, and the audio was recorded at 8 kHz synchronized with the video. All the recorded words in the videos were automatically segmented based on the audio signal, by detecting zero-crossings and energy level in a frame wise basis.

The audio signal is partitioned in frames with the same rate as the video frame rate (60 frames per seconds). For a given frame t, the first eleven non-DC Mel-Cepstral coefficients are computed and used to compose a vector denoted as $\mathbf{a}_t$. In order to take into account the audio-visual co-articulation, the information of t_{c_a} preceding and t_{c_a} subsequent frames is used to form the audio feature vector at frame t, $\mathbf{o}_{at} = \left[\mathbf{a}_{t-t_{c_a}}, \ldots, \mathbf{a}_t, \ldots, \mathbf{a}_{t+t_{c_a}}\right]$.

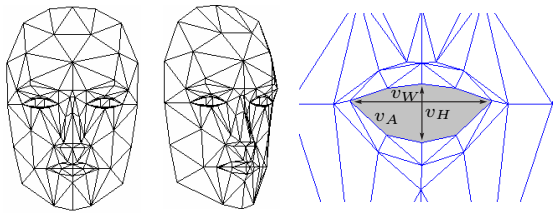
Visual features are represented in terms of a simple 3D face model, namely *Candide-3* [1]. This 3D face model, depicted in Fig. 1(a), has been widely used in computer graphics, computer vision and model-based image-coding applications. The advantage of using the Candide-3 model is that it is a simple generic 3D face model, adaptable to different real faces, that allows to represent facial movements with a small number of parameters. The method proposed by the present authors in [8] is used to extract visual features related to mouth movements during speech. As it is described in [8], this visual information is related to the generic 3D model and it does not depend on the particular face being tracked, *i.e*, this method retrieves normalized mouth movements. The mouth shape at each frame t is then used to compute three visual parameters, *viz.*, mouth height (v_H), mouth width (v_W) and area between lips (v_A), as depicted in Fig. 1(b). These three parameters are used to represent the visual information at frame t, denoted as $\mathbf{v}_t$. Similarly to the case of acoustic information, t_{c_v} preceding and t_{c_v} subsequent frames are used to form the visual feature vector at frame t, $\mathbf{o}_{vt} = \left[\mathbf{v}_{t-t_{c_v}}, \ldots, \mathbf{v}_t, \ldots, \mathbf{v}_{t+t_{c_v}}\right]$.

For a particular frame t, the audio-visual feature vector is composed by the concatenation of the associated acoustic and visual feature vectors, that is

$$\mathbf{o}_{avt} = \left[\mathbf{o}_{at}, \mathbf{o}_{vt}\right]. \tag{1}$$

3 Early Integration

In most applications the acoustic channel is corrupted by noise, degrading the recognition rates of audio-only speech recognition systems. The proposed system aims to improve the recognition rates in these situations, by fusing audio and visual features. In the presence of noise in the acoustic channel, the efficiency of a classifier based on audio-only information decreases as the SNR decreases. On the other hand, the efficiency of a visual information classifier remains constant, since it does not depends on SNR. However, the use of only visual information is usually not enough to obtain relatively good recognition rates. It has been shown in several works in the literature [4][7][6], that the use of audio-visual



(a) (b)

Fig. 1. (a) *Candide-3* face model. (b) Visual parameters.

feature vectors (early integration) improves the recognition rate in the presence of noise in comparison to the audio-only case. In this section, the performances of audio, visual, and audio-visual classifiers are evaluated using audio-visual features extracted from the compiled database, described in section 2. Then, these results are used to derive the proposed late integration strategy described in section 4.

Visual classifier. The visual feature vector $\mathbf{o}_{vt}$ at frame t is composed by the concatenation of the visual information contained in t_{c_v} preceding and t_{c_v} subsequent frames (see section 2). Experiments with 0 to 7 frames of coarticulation (t_{c_v}) were carried out. It must be noted that there is no need to carry out these tests considering different SNRs, since the visual features are not affected by the acoustic noise. The results of these experiments are depicted in Fig. 2, using boxplot representation. As it is customary, the top and bottom of each box are the 75th and 25th percentiles of the samples, respectively, the line inside each box is the sample median, and the notches display the variability of the median between samples. These results were computed across all the words in the vocabulary.

Audio classifier. Similarly to the case of visual feature vectors, the audio feature vector $\mathbf{o}_{at}$ at frame t is composed by the concatenation of the acoustic information contained in t_{c_a} preceding and t_{c_a} subsequent frames. To select the optimum value of t_{c_a}, experiments with 0 to 6 frames of coarticulation were performed. Since the efficiency of the audio classifier depends on the SNR, these experiments were carried out using several SNR levels for two types of noise: additive Gaussian noise and Babble noise. In Fig. 3, the results derived from these experiments are shown, where only the medians for each noise level and coarticulation parameter, are depicted for visual clarity reasons.

Audio-Visual classifier. The audio-visual fusion (early integration) proposed in this paper is based on the concatenation of the audio and visual feature vectors associated to each frame t, as stated in (1). Thus, there are two parameters that define the audio-visual vector: t_{c_a} and t_{c_v}. Modifying these values

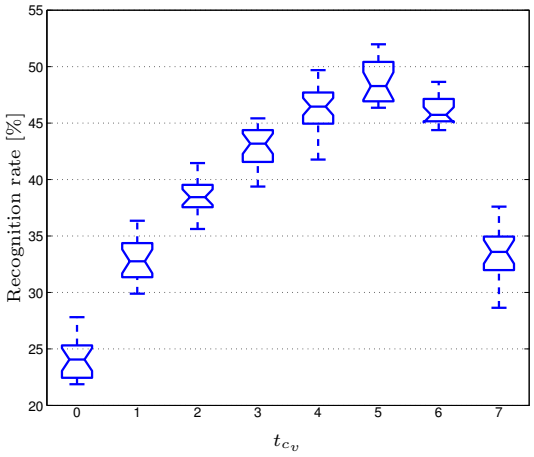


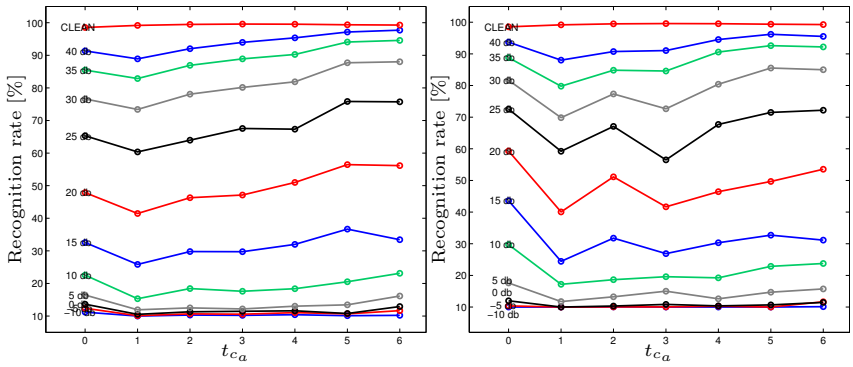
Fig. 2. Recognition rate of the visual classifier using different values of t_{c_v}.

different structures can be obtained. In a similar fashion than for the audio classifier and video classifier, experiments were performed for t_{c_a} and t_{c_v} ranging from 0 to 5. These tests were carried out considering different SNRs for the cases of Gaussian and Babble noises. Figure 4 shows the recognition rates obtained for the different SNRs and the two considered noises, for three particular audio-visual fusion configurations, namely

- $t_{c_a} = 1$ and $t_{c_v} = 5$, denoted as A_1V_5,
- $t_{c_a} = 5$ and $t_{c_v} = 5$, denoted as A_5V_5,
- $t_{c_a} = 5$ and $t_{c_v} = 1$, denoted as A_5V_1.

It can be noted from Fig. 4 that the better performance at low SNRs is obtained for the case of configuration A_1V_5, while configurations A_5V_5 and A_5V_1 present the better performances at high SNRs. The performance of the remaining configurations lies between these curves following the same properties.

Considering the results associated to each classifier, depicted in Figures 2, 3 and 4, it can be clearly noted that the audio classifier performs better than the visual one for high SNRs and viceversa. The combination of audio-visual features leads to an improvement of the recognition rates in comparison to the audio-only case. However, for the case of low SNRs, the audio-visual classifier performs worse than the visual one since fused audio-visual features are degraded by the highly corrupted acoustic data. Using different combination of acoustic and visual features, different performances can be obtained. For instance, if the audio-visual features contains more visual than acoustic information, the performance at low SNRs is improved since visual information is more reliable in this case. However, the efficiency at high SNRs is deteriorated, where the acoustic information is more important. Even for cases where a small portion of



(a) Babble Noise (b) Gaussian noise

Fig. 3. Efficiency of the acoustic classifier using different values of t_{c_a} and different SNRs, for the cases of considering (a) Babble noise, and (b) Gaussian noise.

audio information is considered, a notorious improvement could be obtained for low SNRs, but the efficiency at high SNRs could be worse than for the audio-only case. Thus, there exists a trade-off between performance at low and high SNRs.

4 Late Integration

Taking into account the analysis presented in the previous section, the recognition system proposed in this paper combines three different classifiers based on audio, visual and audio-visual information, respectively, aiming at recognizing the input word and maximizing the efficiency over the different SNRs. In the training stage, a combined classifier is trained for each particular word in the vocabulary. Then, given an audio-visual observation sequence associated to the input word to be recognized, denoted as O_{av}, which can be partitioned into acoustic and visual parts, denoted as O_a and O_v, respectively, the probability (P_i) of the proposed combined classifier corresponding to the i-class is given by

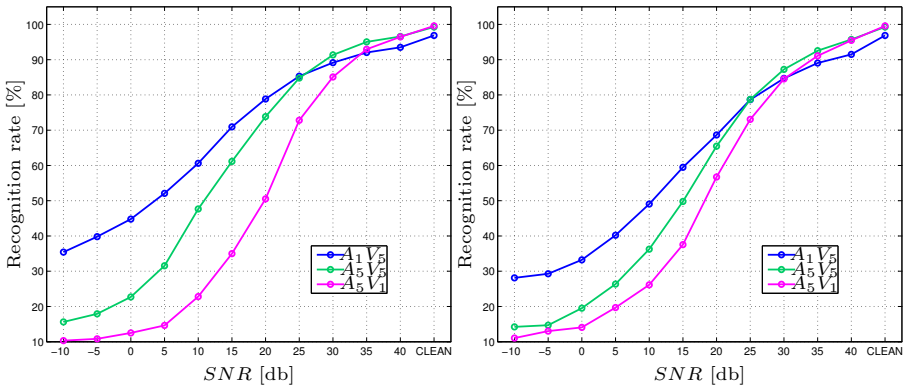
$$P_i = P\left(O_a|\lambda_i^a\right)^{\alpha} P\left(O_v|\lambda_i^v\right)^{\beta} P\left(O_{av}|\lambda_i^{av}\right)^{\gamma}, \tag{2}$$

where $P\left(O_a|\lambda_i^a\right)$, $P\left(O_v|\lambda_i^v\right)$ and $P\left(O_{av}|\lambda_i^{av}\right)$ are the probabilities corresponding to the audio (λ_i^a), visual (λ_i^v) and audio-visual (λ_i^{av}) classifiers, respectively, and α, β and γ are real coefficients that satisfy the following condition

$$\alpha + \beta + \gamma = 1. \tag{3}$$

The visual (λ_v) classifier is more useful at low SNRs (β is predominant), where the acoustic data is highly corrupted by noise, while at medium levels of SNRs, the audio-visual classifier (λ_{av}) retrieves the better decisions (γ is predominant). For high SNR conditions, an audio classifier (λ_a) is employed (α is predominant).

A block diagram representing this computation is depicted in Fig. 5.



(a) Babble Noise (b) Gaussian noise

Fig. 4. Performance of the audio-visual classifier over the SNRs for three different fusion configurations. (a) Babble noise. (b) Gaussian noise.

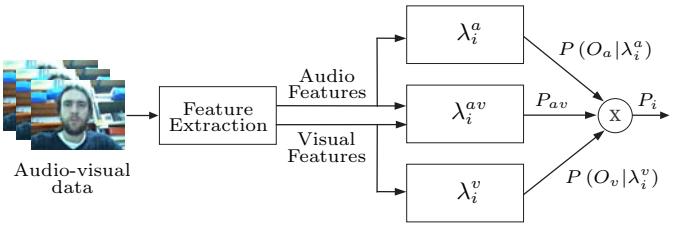


Fig. 5. Schematic representation of the computation of the probability associated to a particular class i for the proposed combined classifier. P_{av} refers to $P(O_{av}|\lambda_i^{av})$.

5 Experimental Results

In this section, the proposed audio-visual speech recognition system is evaluated using the audio-visual database described in section 2. For each experiment reported in this section, 50 round cross-validation was performed, randomly selecting 70% of the database for training and using the remaining 30% for testing. In these experiments the coefficients of the feature vectors were normalized by subtracting the corresponding sample mean and dividing by the corresponding sample variance, computed over the corresponding training set. To evaluate the recognition rates under noisy acoustic conditions, experiments with additive Gaussian noise and Babble noise, with SNRs ranging from -10 dB to 40 dB, were performed.

As it was previously described, the proposed audio-visual speech recognition system combines three classifiers based on audio, visual and audio-visual information, respectively, in order to improve the recognition rates for different SNRs.

These individual classifiers are implemented using left-to-right Hidden Markov Models (HMM) with continuous observations. In order to select the optimum HMM structure, several experiments were performed considering numbers of states in the range from 3 to 7, numbers of Gaussian mixtures from 4 to 11, and full and diagonal covariances matrices. These tests were carried out for the three cases, namely audio, visual and audio-visual features. Based on these experiments, an optimum HMM structure with 4 states, 6 Gaussian mixtures and full covariance matrices was selected for the three different classifiers.

5.1 Classifier selection

For the visual classifier, the results depicted in Fig. 2 shown that the higher accuracy was obtained for 5 frames of coarticulation, which corresponds to a visual feature vector $\mathbf{o}_{vt}$ composed by 33 parameters. In the time domain, this corresponds to a sliding window of 183 msec approximately. Thus, $t_{c_v} = 5$ was adopted for this classifier.

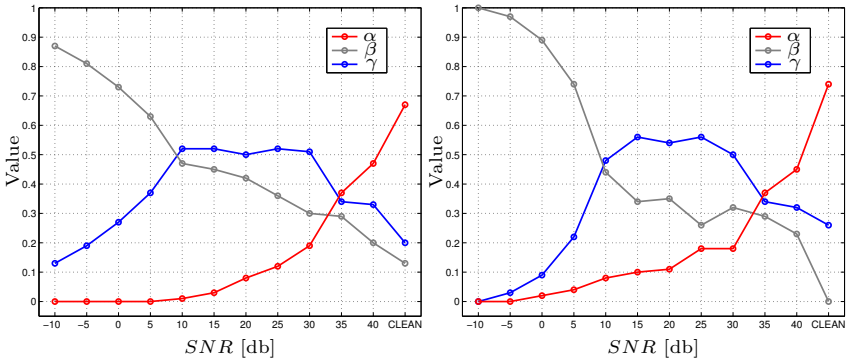
For the audio classifier, it must be noted that the selection of t_{c_a} should be done taking into account that the contribution of this classifier to the final decision stage is important at high SNR conditions. For that reason and looking at Fig. 3, $t_{c_a} = 4$ or $t_{c_a} = 5$ or $t_{c_a} = 6$ could be selected. In order to reduce the dimensionality of the resulting audio feature vectors, without significatively affecting the efficiency of the classifier, $t_{c_a} = 4$ was adopted, which corresponds to audio feature vectors composed by 99 parameters. In the time domain, this corresponds to a sliding window of 150 msec.

Regarding the selection of the optimal audio-visual classifier configuration to be used at the final decision stage, it must be taken into account that the contribution of this classifier is important at low and middle range SNR conditions, since at high SNR the audio classifier provides more accurate decisions. Thus, from Fig. 4 an adequate configuration for this purpose is the one using $t_{c_a} = 1$ and $t_{c_v} = 5$, *i.e.*, configuration A_1V_5.

5.2 Decision level integration

As mentioned in section 4, the combination of the probabilities computed from the independent classifiers, is carried out by the weighted multiplication of the individual probabilities, see Eq. (2), where coefficients α, β and γ modify the contribution to the final decision of the audio, visual and audio-visual classifiers, respectively. The values of these coefficients should be modified for the different SNRs, so that the higher contribution at low SNR comes from the visual classifier, at medium SNRs from the audio-visual classifier, and at high SNRs from the audio classifier. Several experiments were performed using different possible combinations of them to achieve the optimum values. The results of these test are depicted in Fig. 6. For both cases of considering Gaussian and Babble noises, it can be seen that the optimum value of α is the lower one at low SNRs, and it increases as the SNR increases, becoming the higher one at clean audio. On

the other hand, the optimum values of coefficient β present an inverse evolution. While for the case of coefficient γ the higher values are at medium SNRs.



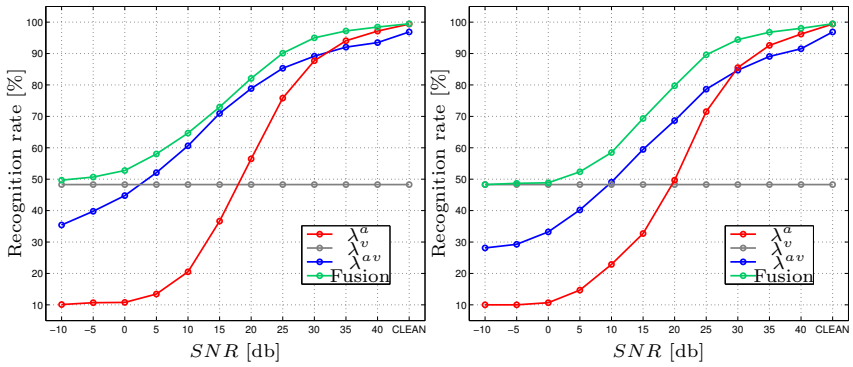
(a) Babble Noise (b) Gaussian noise

Fig. 6. Optimum values for coefficients α, β and γ over the SNRs. (a) Babble noise. (b) Gaussian noise.

In Fig. 7 the obtained recognition rates of the proposed fusing strategy over the SNRs, using the optimum values for the weighting coefficients α, β and γ, are presented. In this figure, the recognition rates corresponding to the audio, visual and audio-visual classifiers are also depicted. It is clear that the proposed objective of improving the recognition rates through the different SNRs has been accomplished.

6 Conclusions

Improvements of speech recognition rates by the incorporation of visual data related to the mouth movements and the late integration of different classifiers are presented in this paper. An isolated Spanish digit recognition system based on audio-visual information was developed to test the proposed system. The acoustic information is represented by mel-frequency cepstral coefficients, while the visual information is represented by coefficients related to mouth shape. Three classifiers based on audio, visual and audio-visual information, respectively, are combined in the proposed system in order to improve the recognition rates through a wide range of signal-to-noise ratios. A Spanish audio-visual database was compiled in order to evaluate the efficiency of the system, considering noisy conditions in the acoustic channel. The experimental results show that a significant improvement is achieved when the visual information is considered. It is important to note that, the absolute recognition rates could be further improved by considering well-known strategies usually employed in speech recognition, for instance, using delta mel-cepstral coefficients in the audio features,



(a) Babble Noise (b) Gaussian noise

Fig. 7. Recognition rates of the proposed fusing strategy, audio, visual and audio-visual classifiers.

including noisy features in the training stage, etc. Work is in progress, where the extension of the proposed system to the case of continuous speech recognition is considered.

References

1. Ahlberg, J.: Candide-3 - an updated parameterised face. Tech. rep., Department of Electrical Engineering, Linkping University, Sweden (2001)
2. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia 2(3), 141–151 (Sep 2000)
3. Estellers, V., Gurban, M., Thiran, J.: On dynamic stream weighting for audio-visual speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 20(4), 1145–1157 (2012)
4. Jaimes, A., Sebe, N.: Multimodal human-computer interaction: A survey. Comput. Vis. Image Understand 108(1-2), 116–134 (2007)
5. Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.: A coupled hmm for audio-visual speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (CASSP02). pp. 2013–2016 (2002)
6. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent advances in the automatic recognition of audio-visual speech. In: PROC. IEEE. vol. 91, pp. 1306–1326 (2003)
7. Shivappa, S., Trivedi, M., Rao, B.: Audiovisual information fusion in human computer interfaces and intelligent environments: A survey. Proceedings of the IEEE 98(10), 1692–1715 (2010)
8. Terissi, L., Gómez, J.: 3D head pose and facial expression tracking using a single camera. Journal of Universal Computer Science 16(6), 903–920 (2010)